12th International Electronic Conference on Synthetic Organic Chemistry (ECSOC-12) 1-30 November 2008
http://www.usc.es/congresos/ecsoc/12/ECSOC12.htm & http://www.mdpi.org/ecsoc-12

[G0011]

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 Arc Repressor

Yovani Marrero-Ponce,[1,2,3*] Sadiel E. Ortega-Broche[1,4] Yunaimy Echevería Díaz,[1] Francisco Torrens,[2] and Facundo Pérez-Giménez.[3]

[1]Unit of Computer-Aided Molecular "*Biosilico*" Discovery and Bioinformatic Research (CAMD-BIR Unit), Faculty of Chemistry-Pharmacy. Central University of Las Villas, Santa Clara, 54830, Villa Clara, Cuba.

[2]Institut Universitari de Ciència Molecular, Universitat de València, Edifici d'Instituts de Paterna, Poligon la Coma s/n, E-46071 Valencia, Spain.

[3]Unidad de Investigación de Diseño de Fármacos y Conectividad Molecular, Departamento de Química Física, Facultad de Farmacia, Universitat de València, Spain.

[4]Department of Physiology, Medical School "Faustino Pérez Hernández", Km # 3 Circumvallation, Sancti-Spíritus, Cuba.

[*]***Corresponding author:***

Fax: 53-42-281130 [or 53-42-281455] (Cuba) and 963543156 (València)
Phone: 53-42-281192 [or 53-42-281473] (Cuba) and 963543156 (València)
Cell: 610028990
***e*-mail**: ymarrero77@yahoo.es; ymponce@gmail.com or yovanimp@uclv.edu.cu
URL: http://www.uv.es/yoma/

ABSTRACT

A new set of amino-acid based bio-macromolecular descriptors support on a bilinear map are presented. This novel approach to bio-macromolecular design from a linear algebra point of view is relevant to protein QSAR/QSPR studies. These biochemical descriptors are based on the computation of bilinear maps on $\Re^n$ [$b_{mk}(\bar{x}_m, \bar{y}_m) : \Re^n \, x \, \Re^n \rightarrow \Re$] in canonical basis. Protein's bilinear indices are calculated from k^{th} power of non-stochastic and stochastic graph–theoretic electronic-contact matrices, M_m^k and ${}^sM_m^k$, respectively. That is to say, the k^{th} non-stochastic and stochastic protein's bilinear indices are calculated using M_m^k and ${}^sM_m^k$ as matrix operators of bilinear transformations. Moreover, biochemical information is codified by using different pair combinations of amino-acid properties as weightings (*z*-values, side-chain isotropic surface area (ISA), amino-acids atomic charges (ECI) and hydrophathy index (Kyte-Doolittle scale; HPI). Quantitative models that discriminate near wild-type stability alanine-mutants from the reduced-stability ones in training and test series were obtained. Non-stochastic and stochastic equations permitted the correct classification of 100% (41/41) and 97.56% (40/41) of proteins in the training set, respectively. Correct classification in test sets were 91.67% for both models. In order to predict Arc alanine-mutant's melting temperature (t_m), lineal regression models were developed. The linear model obtained by using non-stochastic bilinear indices explains almost 84% of the variance of the experimental t_m (R = 0.91 and s = 4.50°C) as long as the stochastic bilinear indices-based equation describe 81% of the t_m variance (R = 0.90 and s = 5.01°C). The Leave-one-our press statistics, evidenced high predictive ability of both models (q^2 = 0.73 and s_{cv} = 4.50 °C for non-stochastic and q^2 = 0.64 and s_{cv} = 5.01 °C for stochastic bilinear indices). Moreover, non-stochastic and stochastic protein's bilinear indices produced rather linear piecewise regressions (R of 0.95 and 0.96, correspondingly) between protein-backbone descriptors and t_m values for alanine-mutants of Arc repressor. Both obtained break-point values were 51.87°C and characterized two mutant's clusters as well as coincided perfectly with the experimental scale. Therefore, we can use the linear discriminant analysis and piecewise models in combination to classify and predict the stability of the mutant Arc homodimers. Protein's bilinear indices models compared favorably with several bio-macromolecular descriptors previously reported. These models also permitted the interpretation of the driving forces of such a folding process, indicating that topologic/topographic protein's backbone interactions control the stability profile of wild-type Arc and its alanine-mutants.

Keywords: Protein Stability, Arc Repressor, Alanine-Substitution Mutant,

TOMOCOMD-CAMPS Software, Bilinear Indices, QSAR, Linear Discriminant

Analysis, Linear Multiple Regression, Piecewise Linear Regression.

Running head: *Predicting Protein Stability of Arc Repressor with Protein Bilinear Indices…*

1. INTRODUCTION

Knowledge of relationship between protein stability and structural features are useful in biotechnology and pharmaceutical industries. Its understanding might allow improving efficiency of synthesis, purification and storage processes so as development of more effective drugs against many diseases. Taking into account the above, we can understand why prediction of protein structure, stability and its specific ligand-binding arise as the main goal of protein science.[1]

Development of genetic and protein engineering facilitate the studies about how structural-changes affect protein functions and the first tentative steps protein design are underway. Moreover, knowledge of factors that determine the stability of a particular protein enables us to find out important features concerning their structure and function. Theoreticians use derived data from protein engineering experiments to benchmark *in silico* calculations that will eventually be used for designing rational changes in protein stability. In fact, predicting protein structures and stability is a fundamental goal in molecular biology nowadays. Even predicting changes in structure and stability induced by point mutations has immediate application in computational protein design.[2-5] Although free energy simulations have accurate predicted relative stabilities of point mutants,[6, 7] the computational cost that the most of the methods actually demand are extremely high to test the large number of mutations studied in protein design applications.

In this context, the computational study of structure/stability relationships has become an important area in protein science. Numerous researchers worldwide have worked out models to predict the stability of mutants of a wild protein. For instance, Shortle et al. have studied 118 mutants of *Staphylococcal* nuclease. Similarly, other researchers have modelled the stability of 145 mutants of T4 Lysozyme, 96 mutants of

Barnase, and 71 mutants of Chymotrypsin in what seem to be the models with the largest mutated proteins. Another important study involved modeling the stability of 66 mutants of GeneV, 65 mutants of Human lysozyme, and 58 mutants of protein L. Other noteworthy studies concerned 40 mutants of Trypsin inhibitor, 38 mutants of TNFn3, and 31 mutants of FKBP12. Models have also been reported for proteins with more than 10 mutants but fewer than 30, such as ACBP, Ribonuclease T1, Ribonuclease H, α-Lactalbumin, Hen Lysozyme, Subtilisin inhibitor, U1A, ISO-1 cytochrome C, and Trp synthase. Other, less-mutated proteins that have been studied include CD2, Calbindin, Apomyoglobin, Adrenodoxin, Cold shock, ribonuclease A and λ-CRO. As summarized in Zhou and Zhou's excellent work, a total of 35 proteins with their respective 1023 mutants have been studied and these include all of the examples outlined above. In their review, Zhou and Zhou not only provide an excellent overview of this field but also use the data from the 1023 mutant stability tests to develop what seems to be one of the largest unified models to date.[8]

Others important approaches for predicting protein stability are based on Force fields by using fast algorithms for protein energy calculations. Examples of such algorithms are the helix/coil transition algorithm AGADIR[9] or FOLDEF, a fast and accurate EEEF (empirical data-based energy function) approach based on AGADIR algorithm that uses a full atomic description of the structure of the proteins reported by Guerois et al. for predicting conformational stability of more than 1000 mutants.[10] Otherwise, Gromiha et al. reported stability prediction studies not based on protein force-field calculations but focused on correlations of free energy change with 3D structure, sequence information and amino acid properties such as hydrophobicity, accessible surface area, etc.[11, 12] Recently, Frenz[13] reported an Artificial Neural Networkbased model for predicting the stability of Staphylococcal Nuclease mutants by

using amino-acid similarity scores as network inputs. Besides, by using a combination of neural network and support vector machine predictors as well as sequence and 3D structure information from a data set of more than 2000 mutants, Capriotti et al.[14, 15] described the change of protein free energy changes upon mutations. More recently, Fernádez et al.[7] reported the used of Bayesian-regularized genetic neural networks for modelling protein conformational stability of Chymotrypsin inhibitor 2 mutants. In this manuscript, the authors redefined a well-know 3D-molecular descriptors (radial distribution function) for small-to-medium organic molecules to codify the 3D molecular structure of proteins. In connection with, is rather important remarkable that although the search of novel molecular descriptors to seek quantitative-structure–activity-relationships QSAR nowadays constitutes a widely covered field with more than 1000 molecular descriptors introduced,[16] the search for newer molecular descriptors for proteins can be classified as an emerging area, being a pioneering work the one on the radius of gyration reported by Flory.[17] More recently, other approaches have been put forward as potential sources for successful biopolymer descriptors, such as Roy et al.,[18] Casanovas et al.,[19] and Leong and Mogenthaler representations;[20] Arteca's average over crossing number,[21] Randic's band average widths,[22] the sequence-order-coupling numbers,[23] folding degree index (I_3)[24] Kyle–Doolittle hydrophobicity,[25] and so on. In this sense, the approach of redefinition of old molecular descriptors, that has been successfully used for long-time ago in QSAR/QSPR studies, into new protein indices is a good alternative and an active field in protein science. In this context, some specific and very successful indices (for small molecules) that use the concept of Shannon's entropy from the point of view of the information theory have proven to be very effective in drug design[26, 27] and has been recently redefined in order to characterized the molecular structure of proteins.[28-30]

In these studies, the researchers attempt to extend your method to encompass protein stability studies—specifically how alanine substitution mutation on *Arc* repressor wildtype protein affects protein stability–by means of Linear Discriminant Analysis (LDA). The *Arc repressor* protein provides an attractive system in which to address this issue because it is small (53 AAs), and amenable to genetic and biophysical studies[31-36] This is a homodimer protein with a globular domain formed by the intertwining of their monomers. Its secondary structure consists on two anti-parallel β-sheets from residues 8-14, and α-helices formed by residues 15-30 and 32-48.[31-36]

Recently, one of present author, M-P, Y., also redefined two molecular descriptors families, namely atom- and bond-based quadratic and linear indices,[37-43] like bio-macromolecular indices.[44-47] In these studies, our group proposed a novel extended method to represent and to codify (translation of molecular structure into numerical parameters) the molecular structure of proteins and nucleic acids, in which each amino-acid residues can be depicted using a lower level representation, that is, a pseudo-atoms rather than by an all-atom representation. This approach has been successfully employed in bioinformatics studies showing promising results in the modeling of the interaction between drugs and HIV packaging-region RNA and Arc repressor stability.[44-47] Newly, the present author purposed a novel algebraic algorithm like a extended and generalized form of precedent bond- and atom-based molecular indices for small-to-medium organic-chemicals, namely *global (total) and locals (i.e., atom- or bond-type, group-type, etc) molecular bilinear indices*.[48-52] These new molecular descriptors based on the linear algebra theory (bilinear map) and discrete mathematics (graph-theory) describes changes in the electron distribution with time throughout the molecular backbone (graph–theoretic electronic–structure models) and the complete model (application) can be seen as an intermediate between the quantitative quantum-

mechanical Schrödinger equation and classical chemical bonding ideas.[53] It has been successfully employed in QSPR/QSAR studies,[48-52] such as: a) The fast-track experimental discovery of novel tyrosinase inhibitors drug-like compounds,[49, 51] b) biosilico discovery by using virtual screening of new trichomonacidals,[52] c) Codification of chirality and other 3D structural features constitutes,[50] d) others application in course. Finally, but not less, the molecular bilinear indices has shown better behaviour in the description and prediction the several properties/activities than their counterparts quadratic or linear maps as well as with others 0D-3D molecular descriptors.[48-52]

Therefore, the main purpose of the current paper is to present new extended sets of bio-macromolecular descriptors, namely *non-stochastic and stochastic protein bilinear indices* and establish their abilities (both total and local) for the description of the macromolecular structure by predicting protein stability effects of a complete set of Alanine substitutions in Arc repressor. This study also permit to compare our novel approach to others reported up to now for this *in silico* experiment.

2. MATHEMATICAL DEFINITION

In previous reports, we outline outstanding features concerned with the theory of 2D atom-based *TOMOCOMD-CARDD* MDs. This method codifies the molecular structure by means of mathematical quadratic, linear and bilinear transformations.[37-43, 48-52] In order to calculate these algebraic maps for a molecule, the atom-based molecular vector, $\bar{x}$ (vector representation) and k^{th} "non-stochastic and stochastic graph–theoretic electronic-density matrices", $\mathbf{M}^{k}$ and $\mathbf{S}^{k}$ correspondingly (matrix representations), are constructed. In connection, atom-based quadratic and linear indices were recently extended to structural codification and biological properties prediction of

biopolymers (proteins and nucleic acid) using aminoacid-adjacency relationships and chemical-information codification.[44-47] Therefore the structure of this section will be as follows: 1) a background in aminoacid-based macromolecular vector and non-stochastic and stochastic graph–theoretic electronic-contact matrices will be described in the next subsections (2.1 and 2.2, respectively), and 2) an outline of the mathematical definition of bilinear maps and a definition of our procedures will be develop in subsections 2.3 and 2.4, correspondingly.

2.1. Chemical Information and Aminoacid-based Macromolecular Vector

In analogy to the molecular vector $\bar{x}$ used to represent organic molecules[54] [37-43, 48-52] we introduce here the macromolecular vector ($\bar{x}_m$). The components of this vector are numeric values, which represent a certain side-chain amino-acid property. These properties characterize each kind of amino-acid (R group) within a protein. Such properties can be z-values,[55] side-chain isotropic surface area (ISA) and atomic charges (ECI) of the amino-acid,[56] hydropathy index (Kyte-Doolittle scale; HPI)[57] as well as other hydrophobicity scales such as Hopp-Woods [58], and so on. For instance, the $z_{1(AA)}$ scale of the amino-acid AA takes the values $z_{1(V)}$ = -2.69 for valine, $z_{1(A)}$ = 0.07 for alanine, $z_{1(M)}$ = 2.49 for methionine and so on.[55, 56] Table 1 depicts several side-chain descriptors for the natural amino-acids.[55-57]

Table 1 comes about here (see end of the document)

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 $\Re^5$, $\Re^{10}$, $\Re^{15}$,..., $\Re^n$, respectively. Where *n* is the dimension of the real sets ($\Re^n$).

This approach allows us encoding peptides such as SKEERN through out the macromolecular $\bar{x}_m = [1.96 \quad 2.84 \quad 3.08 \quad 3.08 \quad 2.88 \quad 3.22]$, in the z_1-scale (see Table 1 for

more details). This vector belongs to the product space $\Re^6$. The use of other scales defines alternative macromolecular vectors.

Now, if we are interested to codify the chemical information by means of two different macromolecular vectors, for instance, $\bar{x}_m = [x_{m1},\ldots, x_{mn}]$ and $\bar{y}_m = [y_{m1},\ldots, y_{mn}]$; then different combinations of macromolecular vectors ($\bar{x}_m \neq \bar{y}_m$) are possible when a weighting scheme is used. In the present report, we characterized each amino-acid with the biochemical parameters shown in Table 1. From this weighting scheme, fifteen (or thirty if $\bar{x}_{m\,w}$- $\bar{y}_{m\,z} \neq \bar{x}_{m\,z}$- $\bar{y}_{m\,w}$) combinations (pairs) of macromolecular vectors ($\bar{x}_m, \bar{y}_m; \bar{x}_m \neq \bar{y}_m$) can be computed, $\bar{x}_{m\,z1}$- $\bar{y}_{m\,z2}$, $\bar{x}_{m\,z1}$- $\bar{y}_{m\,z3}$, $\bar{x}_{m\,z1}$- $\bar{y}_{m\,HPI}$, $\bar{x}_{m\,z1}$- $\bar{y}_{m\,ISA}$, $\bar{x}_{m\,z1}$- $\bar{y}_{m\,ECI}$, $\bar{x}_{m\,z2}$- $\bar{y}_{m\,z3}$, $\bar{x}_{m\,z2}$- $\bar{y}_{m\,HPI}$, $\bar{x}_{m\,z2}$- $\bar{y}_{m\,ISA}$, $\bar{x}_{m\,z2}$- $\bar{y}_{m\,ECI}$, $\bar{x}_{m\,z3}$- $\bar{y}_{m\,HPI}$, $\bar{x}_{m\,z3}$- $\bar{y}_{m\,ISA}$, $\bar{x}_{m\,z3}$- $\bar{y}_{m\,ECI}$, $\bar{x}_{m\,HPI}$- $\bar{y}_{m\,ECI}$, $\bar{x}_{m\,HPI}$- $\bar{y}_{m\,ECI}$ and $\bar{x}_{m\,ISA}$- $\bar{y}_{m\,ECI}$. Here, we used the symbols $\bar{x}_{m\,w}$- $\bar{y}_{m\,z}$, where the subscripts w and z mean two amino-acid properties from our weighting scheme and a hyphen (-) expresses the combination (pair) of two selected aminoacid-label biochemical properties.

In order to illustrate this, let us consider the same peptide as in the example above SKEERN and the following weighting scheme: z_1 and z_2 ($\bar{x}_{m\,z1}$- $\bar{y}_{m\,z2}$ = $\bar{x}_{m\,z2}$- $\bar{y}_{m\,z1}$). The following macromolecular vectors $\bar{x}_m$ = [1.96 2.84 3.08 3.08 2.88 3.22] and $\bar{y}_m$ = [-1.63 1.41 0.39 0.39 2.52 1.45] are obtained when we use z_1 and z_2 as chemical weights for codifying each amino-acid in the example peptide in $\bar{x}_m$ and $\bar{y}_m$ vectors, respectively (see also Table 2).

Table 2 comes about here (see end of the document)

2.2. Background in non-Stochastic and Stochastic Graph–Theoretic Electronic-Contact Matrices.

In molecular topology, molecular structure is expressed, generally, by the hydrogen-suppressed graph. That is, a molecule is represented by a graph. Informally a graph G is a collection of vertices (points) and edges (lines or bonds) connecting these vertices.[59-61] In more formal terms, a simple graph G is defined as an ordered pair [V(G), E(G)] which consists of a nonempty set of vertices V(G) and a set E(G) of unordered pairs of elements of V(G), called edges.[59-61] In this particular case we are not dealing with a simple graph but with a so-called pseudograph (G). Informally, a pseudograph is a graph with multiple edges or loops between the same vertices or the same vertex. Formally: a pseudograph is a set V of vertices along a set E of edges, and a function *f* from E to $\{\{u,v\} | u,v \text{ in V}\}$ (The function *f* shows which pair of vertices are connected by which edge). An edge is a loop if $f(e) = \{u\}$ for some vertex *u* in V.[37, 62, 63]

In the other hand, Anfinsen's experiments with small proteins demonstrated that protein amino-acid sequence encodes their peptidic backbone folding. However, nowadays, the merely knowledge about amino-acid sequence of a protein don't provide us its three-dimensional structure. Primary structure of proteins consists in unbranched amino-acid sequences, linked by amide bonds between the α-carboxyl group of one residue and the α-amino group of the next. Three-dimensional distribution of all atoms in a protein is referred to as the protein's tertiary structure. Whereas the term secondary structure refers to the spatial arrangement of amino-acid residues that are adjacent in the primary structure, tertiary structure includes longer-range aspects of amino-acid sequence. Last, individual polypeptidic chains into multi-subunit proteins are organized in three-dimensional complexes reaching quaternary-structural levels. As previously outlined, essential information for proteins folding are contained in amino-acid sequence, more specifically in amino-acid side-chains of polypeptidic chain.

Taken into account the above statement, in this paper we develop a graph-theoretical model to represent the molecular structure of proteins. This is called macromolecular graph. Here, graph's vertices are C_α-atoms into polypeptide backbone and edges are both covalent interactions between amino-acids (peptidic bonds) and non-covalent interactions between amino-acid side-chains in same or different subunit. Non-covalent interactions can happen too between an amino-acid side-chain and its main-chain, then this amino-acid represent a pseudo-vertice into macromolecular pseudograph. These interactions can be considerer like contacts, which can be among amino-acid near of far in the polypeptide backbone, that is, the contact can be subdivided in to short-, medium- and large-contacts. Table 2 displays how to depict two interacting polypeptidic chains by means a macromolecular pseudograph because the hetero-dimer (SKEERN) contains an amino-acid having hydrogen bond between its side-chain and its main-chain atom.

The nxn k^{th} non-stochastic graph–theoretic electronic-contact matrix, M_m^k, is a square and symmetric matrix, where n is the number of amino-acids in the protein.[44, 47] The coefficients ${}^k m_{ij}$ are the elements of the k^{th} power of M_m and are defined as follows:

m_{ij} = 1 if $i \neq j$ and $\exists\, e_k \in E(G_m)$ (**1**)

= 1 if $i = j$ and the amino-acid i has a hydrogen-bond between its side-chain and its main-chain atom.

= 0 otherwise

where $E(G_m)$ represents the set of edges of G_m.

The matrix M_m^k provides the numbers of walks of length k that links every pair of vertices v_i and v_j. For this reason, each edge in M_m^1 represents a peptidic bond (covalent

bond) or a hydrogen-bond as well as salt bridge interaction (non-covalent bond) between amino-acids *i* and *j*.

On the other hand, the k^{th} stochastic graph–theoretic electronic-contact matrix of G_m, ${}^{s}M_m^k$, can be directly obtained from M_m^k. Here, ${}^{s}M_m^k = [{}^{k}sm_{ij}]$, is a square matrix of order *n* (*n* = number of C_α atoms) and the elements ${}^{k}sm_{ij}$ are defined as follows:

$$ {}^{k}sm_{ij} = \frac{{}^{k}m_{ij}}{{}^{k}SUM_i} = \frac{{}^{k}m_{ij}}{{}^{k}\delta_i} \qquad (\mathbf{2}) $$

where, ${}^{k}m_{ij}$ are the elements of the k^{th} power of M_m^k and the SUM of the i^{th} row of M_m^k are named the *k*-order vertex degree of C_α atom *i*, ${}^{k}\delta_i$. It should be remarked that the matrix ${}^{s}M_m^k$ in Eq. **2** has the property that *the sum of the elements in each row* is 1. An *n*x*n* matrix with nonnegative entries having this property is called a "stochastic matrix".[64] Table 3 show the zero, first and second powers of the total non-stochastic and stochastic graph–theoretic electronic-contact matrices of macromolecular pseudograph depicted in Table 2.

Table 3 comes about here (see end of the document)

2.3. Mathematical Bilinear Forms: A Theoretical Framework

In mathematics, a **bilinear form** in a real vector space is a mapping $b : VxV \to \Re$, which is linear in both arguments.[65-70] That is, this function satisfies the following axioms for any scalar α and any choice of vectors $\overline{v}, \overline{w}, \overline{v}_1, \overline{v}_2, \overline{w}_1$ and $\overline{w}_2$.

i. $b(\alpha\overline{v}, \overline{w}) = b(\overline{v}, \alpha\overline{w}) = \alpha b(\overline{v}, \overline{w})$

ii. $b(\overline{v}_1 + \overline{v}_2, \overline{w}) = b(\overline{v}_1, \overline{w}) + b(\overline{v}_2, \overline{w})$

iii. $b(\overline{v}, \overline{w}_1 + \overline{w}_2) = b(\overline{v}, \overline{w}_1) + b(\overline{v}, \overline{w}_2)$

That is, *b* is *bilinear* if it is linear in each parameter, taken separately.

Let V be a real vector space in $\Re^n$ ($V \in \Re^n$) and consider that the following vector set, $\{\bar{e}_1, \bar{e}_2, ..., \bar{e}_n\}$ is a basis set of $\Re^n$. This basis set permits us to write in unambiguous form any vectors $\bar{x}$ and $\bar{y}$ of V, where $(x^1, x^2, ..., x^n) \in \Re^n$ and $(y^1, y^2, ..., y^n) \in \Re^n$ are the coordinates of the vectors $\bar{x}$ and $\bar{y}$, respectively. That is to say,

$$\bar{x} = \sum_{i=1}^{n} x^i \bar{e}_i \tag{3}$$

and,

$$\bar{y} = \sum_{j=1}^{n} y^j \bar{e}_j \tag{4}$$

Subsequently,

$$b(\bar{x}, \bar{y}) = b(x^i \bar{e}_i, y^j \bar{e}_j) = x^i y^j b(\bar{e}_i, \bar{e}_j) \tag{5}$$

if we take the a_{ij} as the nxn scalars $b(\bar{e}_i, \bar{e}_j)$. That is,

$$a_{ij} = b(\bar{e}_i, \bar{e}_j), \text{ to } i = 1,2,\ldots,n \text{ and } j = 1,2,\ldots,n \tag{6}$$

Then,

$$b(\bar{x}, \bar{y}) = \sum_{i,j}^{n} a_{ij} x^i y^j = [X]^T A [Y] = \begin{bmatrix} x^1 & \ldots & x^n \end{bmatrix} \begin{bmatrix} a_{11} & \ldots & a_{jn} \\ \ldots & \ldots & \ldots \\ a_{n1} & \ldots & a_{nn} \end{bmatrix} \begin{bmatrix} y^1 \\ \vdots \\ y^n \end{bmatrix} \tag{7}$$

As it can be seen, the defined equation for b may be written as the single matrix equation (see Eq. **7**), where [Y] is a column vector (an nx1 matrix) of the coordinates of $\bar{y}$ in a basis set of $\Re^n$, and $[X]^T$ (a 1xn matrix) is the transpose of [X], where [X] is a column vector (an nx1 matrix) of the coordinates of $\bar{x}$ in the same basis of $\Re^n$.

Finally, we introduce the formal definition of **symmetric bilinear form.** Let V be a real vector space and b be a bilinear function in VxV**.** The bilinear function b is called symmetric if $b(\bar{x}, \bar{y}) = b(\bar{y}, \bar{x}), \forall \bar{x}, \bar{y} \in V$.[65-70] Then,

$$b(\bar{x},\bar{y}) = \sum_{i,j}^{n} a_{ij} x^i y^j = \sum_{i,j}^{n} a_{ji} x^j y^i = b(\bar{y},\bar{x}) \quad (8)$$

2.4. Non-Stochastic and Stochastic Amino Acid-Based Bilinear Indices: Total (Global) Definition.

The k^{th} non-stochastic and stochastic bilinear indices for a protein, $b_{mk}(\bar{x}_m,\bar{y}_m)$ and ${}^s b_{mk}(\bar{x}_m,\bar{y}_m)$, are computed from these k^{th} non-stochastic and stochastic graph–theoretic electronic-contact matrix, M_m^k and ${}^s M_m^k$ as shown in Eqs. **9** and **10**, respectively:

$$b_{mk}(\bar{x}_m,\bar{y}_m) = \sum_{i=1}^{n} \sum_{j=1}^{n} {}^k m_{ij} x_m{}^i y_m{}^j \quad (9)$$

$${}^s b_{mk}(\bar{x}_m,\bar{y}_m) = \sum_{i=1}^{n} \sum_{j=1}^{n} {}^k sm_{ij} x_m{}^i y_m{}^j \quad (10)$$

where n is the number of amino-acids (C_α atom) in the protein, and $x_m{}^1,\ldots,x_m{}^n$ and $y_m{}^1,\ldots,y_m{}^n$ are the coordinates or components of the macromolecular vectors $\bar{x}_m$ and $\bar{y}_m$ in a canonical basis set of $\Re^n$.

The defined equations (**9**) and (**10**) for $b_{mk}(\bar{x}_m,\bar{y}_m)$ and ${}^s b_{mk}(\bar{x}_m,\bar{y}_m)$ may be also written as the single matrix equations:

$$b_{mk}(\bar{x}_m,\bar{y}_m) = [X_m]^T M_m^{\mathbf{k}} [Y_m] \quad (11)$$

$${}^s b_{mk}(\bar{x}_m,\bar{y}_m) = [X_m]^T {}^s M_m^{\mathbf{k}} [Y_m] \quad (12)$$

where $[Y_m]$ is a column vector (an $nx1$ matrix) of the coordinates of $\bar{y}_m$ in the canonical basis set of $\Re^n$, and $[X_m]^T$ is the transpose of $[X_m]$, where $[X_m]$ is a column vector (an $nx1$ matrix) of the coordinates of $\bar{x}_m$ in the canonical basis of $\Re^n$. Therefore, if we use the canonical basis set, the coordinates $[(x_m{}^1,\ldots,x_m{}^n)$ and $(y_m{}^1,\ldots,y_m{}^n)]$ of any macromolecular vectors ($\bar{x}_m$ and $\bar{y}_m$) coincide with the components of those vectors

[($x_{m1},\ldots,x_{mn}$) and ($y_{m1},\ldots,y_{mn}$)]. For that reason, those coordinates can be considered as weights (R-group in C_α atom, that is to say "amino-acid labels") of the vertices of G_m , due to the fact that components of the molecular vectors are values of some amino-acid property that characterizes each kind of R-chain in protein. The calculation of the three first values of bilinear indices for example protein (see Table 2 and 3) is shown in Table 4.

Table 4 comes about here (see end of the document)

It should be remarked that non-stochastic and stochastic bilinear indices are symmetric and non-symmetric bilinear forms, respectively. Therefore, if in the following weighting scheme, W and Z are used as amino-acid weights to compute these protein's bilinear indices, two different sets of stochastic bilinear indices, ${}^{W\text{-}Z}{}^{s}b_{mk}(\bar{x}_m,\bar{y}_m)$ and ${}^{Z\text{-}W}{}^{s}b_{mk}(\bar{x}_m,\bar{y}_m)$ [because $\bar{x}_{m\,W}\text{-}\bar{y}_{m\,Z} \neq \bar{x}_{m\,Z}\text{-}\bar{y}_{m\,W}$] can be obtained and only one group of non-stochastic bilinear indices ${}^{W\text{-}Z}b_{mk}(\bar{x}_m,\bar{y}_m) = {}^{Z\text{-}W}b_{mk}(\bar{x}_m,\bar{y}_m)$ because in this case $\bar{x}_{m\,W}\text{-}\bar{y}_{m\,Z} = \bar{x}_{m\,Z}\text{-}\bar{y}_{m\,W}$ can be calculated.

2.5. Non-Stochastic and Stochastic Local Bilinear Indices: Amino-acid, Aminoacid-type and Peptide Fragment Bilinear Indices Definition.

In the last decade, Randić[71] proposed a list of desirable attributes for a molecular descriptor. Therefore, this list can be considered as a methodological guide for the development of new topological indices. One of the most important criteria is the possibility of defining the descriptors locally. This attribute refers to the fact that the index could be calculated for the molecule (protein) as a whole but also over certain fragments of the structure itself.

Therefore, in addition to *total bilinear indices* computed for the whole protein, a local-fragment (peptide fragment) formalism can be developed. These descriptors are

termed *local non-stochastic and stochastic bilinear indices*, $b_{mkL}(\bar{x}_m, \bar{y}_m)$ and $^{s}b_{mkL}(\bar{x}_m, \bar{y}_m)$, respectively. The definition of these descriptors is as follows:

$$b_{mk\ L}(\bar{x}_m, \bar{y}_m) = \sum_{i=1}^{n} \sum_{j=1}^{n} {}^{k}m_{ij\ L}\, x_m^{\ i}\, y_m^{\ j} \qquad (\mathbf{13})$$

$$^{s}b_{mk\ L}(\bar{x}_m, \bar{y}_m) = \sum_{i=1}^{n} \sum_{j=1}^{n} {}^{k}sm_{ij\ L}\, x_m^{\ i}\, y_m^{\ j} \qquad (\mathbf{14})$$

where $^{k}m_{ijL}$ [$^{k}sm_{ijL}$] is the k^{th} element of the row "*i*" and column "*j*" of the local matrix M^{k}_{mL} [$^{s}M^{k}_{mL}$]. This matrix is extracted from the M^{k}_{m} [$^{s}M^{k}_{m}$] matrix and contains information referred to the vertices of the specific protein fragments (F_r) and also of the molecular environment in *k* step. The matrix M^{k}_{mL} [$^{s}M^{k}_{mL}$] with elements $^{k}m_{ijL}$ [$^{k}sm_{ijL}$] is defined as follows (see Table 5):

$^{k}m_{ijL}$ [$^{k}sm_{ijL}$] $= {}^{k}m_{ij}$ [$^{k}sm_{ijL}$] *if* both v_i and v_j are vertices (amino-acid) contained within the F_r

$= {}^{1}/_{2}\, {}^{k}m_{ij}$ [$^{k}sm_{ijL}$] *if* v_i or v_j are vertices contained within F_r but not both

= 0 otherwise (**15**)

Table 5 comes about here (see end of the document)

These local analogues can also be expressed in matrix form by the expressions:

$$b_{mkL}(\bar{x}_m, \bar{y}_m) = [X_m]^T M_m^{\mathbf{k}}{}_{\mathbf{L}} [Y_m] \qquad (\mathbf{16})$$

$$^{s}b_{mk}(\bar{x}_m, \bar{y}_m) = [X_m]^T {}^{s}M_m^{\mathbf{k}}{}_{\mathbf{L}} [Y_m] \qquad (\mathbf{17})$$

It should be remarked that the scheme above follows the spirit of a Mulliken population analysis.[72] It should be also pointed out that for every partitioning of a protein into Z macromolecular fragments there will be Z local macromolecular fragment matrices. In this case, if a protein is partitioned into Z molecular fragments, the matrix M^{k}_{m} [$^{s}M^{k}_{m}$] can be correspondingly partitioned into Z local matrices M^{k}_{mL} [$^{s}M^{k}_{mL}$], L =

1,..., Z, and the k^{th} power of matrix M_m^k [${}^sM_m^k$] is exactly the sum of the k^{th} power of the local Z matrices. In this way, the total non-stochastic and stochastic bilinear indices are the sum of the non-stochastic and stochastic bilinear indices, respectively, of the Z macromolecular fragments:

$$b_m(\bar{x}_m, \bar{y}_m) = \sum_{L=1}^{Z} b_{mkL}(\bar{x}_m, \bar{y}_m) \quad \textbf{(18)}$$

$${}^sb_m(\bar{x}_m, \bar{y}_m) = \sum_{L=1}^{Z} {}^sb_{mkL}(\bar{x}_m, \bar{y}_m) \quad \textbf{(19)}$$

In addition, the aminoacid-type bilinear indices can also be calculated. Aminoacid and aminoacid-type bilinear indices are specific cases of local protein bilinear indices. In this sense, the k^{th} amino-acid bilinear indices are calculated by summing the k^{th} amino-acid bilinear indices of all amino-acids of the same amino-acid type in the protein. In the aminoacid-type bilinear indices formalism, each amino-acid in the molecule is classified into an aminoacid-type (fragment), such as apolar, polar uncharged, polar charged, positive charged, negative charged, aromatic, and so on. For all data sets, including those with a common molecular scaffold as well as those with very diverse structure, the k^{th} aminoacid-type bilinear indices provide important information. The calculation of the three first values of local (amino-acid) bilinear indices for example protein (see also Tables 2 and 3) is shown in Table 6.

Table 6 comes about here (see end of the document)

Any local protein's bilinear index has a particular meaning, especially for the first values of *k*, where the information about the structure of the fragment F_R is contained. Higher values of *k* relate to the environment information of the fragment F_R considered within the macromolecular pseudograph.

In any case, a complete series of indices performs a specific characterization of the chemical structure. The generalization of the matrices and descriptors to "superior

analogues" is necessary for the evaluation of situations where only one descriptor is unable to bring a good structural characterization.[16, 71] The local macromolecular indices can also be used together with total ones as variables for QSAR/QSPR modelling of properties or activities that depend more on a region or a fragment than on the macromolecule as a whole.

3. MATERIAL AND METHODS

3.1. Computational Strategies

TOMOCOMD is an interactive program for molecular design and bioinformatics research.[73] The program is composed by four subprograms, each one of them dealing with drawing structures (drawing mode) and calculating 2D and 3D molecular and bio-macromolecular descriptors (calculation mode). The modules are named CARDD (Computed-Aided 'Rational' Drug Design), CAMPS (Computed-Aided Modelling in Protein Science), CANAR (Computed-Aided Nucleic Acid Research) and CABPD (Computed-Aided Bio-Polymers Docking). In this paper we outline salient features concerned with only one of these subprograms: CAMPS. This subprogram was developed based on a user-friendly philosophy without *prior* knowledge of programming skills.

The calculation of total and local macromolecular bilinear indices for any peptide or protein was implemented in the ***TOMOCOMD-CAMPS*** software.[73] The main steps for the application of this method in QSAR/QSPR can be briefly resumed as follows:

1. Draw the macromolecular pseudographs for each protein of the data set, using the software's drawing mode. This procedure is carried out by a selection of the active amino-acid symbol belonging to 'natural' amino-acid code. Here, we

consider covalent (peptidic bond) and non-covalent [hydrogen-bond and other electrostatic interaction (within a chain as well as between chains)] interaction. Afterward, we draw the mutants by changing an AA for alanine and considering that this change only affect the possibility of this region of the protein to form polar interaction (because we suppressed the hydrogen interaction if the former AA had it).

2. Use appropriated amino-acid weights in order to differentiate the side-chain of each amino-acid. In this work, we used as amino-acid property some descriptors for the natural amino-acid: the three z-values,[55] Kyte-Doolittle's hydrophobicity scale,[57] ISA and ECI.[56]
3. Compute the non-stochastic and stochastic protein bilinear indices. They can be performed in the software calculation mode, in which one can select the side-chain properties and the family descriptor previously to calculate the bio-macromolecular indices. This software generates a table in which the rows and columns correspond to the compounds and the $b_{mk}(\bar{x}_m, \bar{y}_m)$, respectively.
4. Find a QSPR/QSAR equation by using statistical techniques, such as multilinear regression analysis (MRA), Neural Networks, Linear Discrimination Analysis (LDA), and so on. That is to say, we can find a quantitative relation between a property ***P*** and the $b_{mk}(\bar{x}_m, \bar{y}_m)$ having, for instance, the following appearance,

$$\boldsymbol{P} = a_0\, b_{m0}(\bar{x}_m, \bar{y}_m) + a_1 b_{m1}(\bar{x}_m, \bar{y}_m) + a_2 b_{m2}(\bar{x}_m, \bar{y}_m) + \ldots + a_k b_{mk}(\bar{x}_m, \bar{y}_m) + \mathrm{c} \quad (\mathbf{20})$$

where ***P*** is the measurement of the property, $b_{mk}(\bar{x}_m, \bar{y}_m)$ [or $b_{mkL}(\bar{x}_m, \bar{y}_m)$] is the k^{th} total [or local] macromolecular non-stochastic bilinear indices, an the $\boldsymbol{a}_k$*'s* are the coefficients obtained by the statistical analysis.
5. Test the robustness and predictive power of the QSPR/QSAR equation by using internal and external cross-validation techniques.

6. Develop a structural interpretation of the obtained QSAR/QSPR model using macromolecular bilinear indices as molecular descriptors.

3.2. Data Sets

Arc is a homodimer in which each monomer intertwines with the other to form a single, globular domain with a well-defined core. Several side-chain hydrogen bond and salt-bridge interactions are involved in the Arc crystal structure. An exhaustive representation of these interactions can be observed in some detail elsewhere (see Fig 1*b* in Reference 34). Nevertheless, an overview of these electrostatic interactions in Arc repressor structure will be given. Hydrogen-bond interactions take place:[34]

i) Between side chain in the same subunit (N29-E36) and; those between side chains in different subunits (R40-S44).

ii) Between a side chain and main-chain atom intersubunit (W14-N34, N34-R13) and; those between a side chain and main-chain atom intrasubunits (E17-E17, S32-S35, S44-R40).

On the other hand, salt-bridge interactions take place:[34]

iii) Between side chain in the same subunit (R16-D20, D20-R23, R31-E36, E36-R40, E43-K46, E43-K47) and; those between side chains in different subunits (E28-R50, R40-E48).

The data of Arc repressor mutant was taken from the literature.[34] In this paper**,** Alanine substitutions were constructed at each of the 51 non-alanine positions in the wild-type Arc sequence. To avoid intracellular proteolysis and purification difficulties, these authors constructed the alanine substitution mutant in backgrounds containing the carboxy-terminal extensions $(His)_6$ (designated st6) or $(His)_6$-Lys-Asn-Gln-His-Glu (designated st11).[74, 75] These tail sequences allow affinity purification, reduce degradation and cause no significant changes in protein stability.[33]

Milla et al.[34] subjected each purified mutant of Arc to thermal and urea denaturation experiments. Stability of the proteins was checked by melting temperature (t_m). The values of t_m for 53 Arc homodimers reported by these authors are given in Tables 7 and 8.

Table 7 and 8 comes about here (see end of the document)

In equilibrium and kinetic unfolding-refolding studies only native Arc dimers and denatured monomers are significantly populated. Thus, folding and dimerization are concerted processes.[31, 34, 35] For this reason, it is important to remember that t_m refers to unfolding of the Arc homodimer. Then, one must take into consideration that each single mutation changes two side-chains in the Arc dimer, being stability effects roughly twice these observed for monomeric proteins. Moreover, changes in stability may arise due to mutation disrupts of a native interaction, when the native structure of the mutant undergoes relaxation, or because of the change on the properties of the denatured mutant protein.[34, 76-79]

3.3. Chemometric Analysis: Classification- and Regression-Based QSAR Model.

Linear Discrimination Analysis (LDA), Linear Multiple Regression (LMR) and the non-linear estimation analysis, Piecewise Linear Regression (PLR) were used to obtain mathematical models. These statistical analyses were carried out with the STATISTICA software package.[80] Forward stepwise was fixed as the strategy for variable selection in the case of LDA and LMR analysis. The tolerance parameter (proportion of variance that is unique to the respective variable) used was the default value for minimum acceptable tolerance, which is 0.01.

LDA is used in order to generate the classifier function on the basis of the simplicity of the method.[81] To test the quality of the discriminant functions derived we used the Wilks' λ and the Mahalanobis distance. The Wilks' λ statistic for overall

discrimination can take values in the range of 0 (perfect discrimination) to 1 (no discrimination). The Mahalanobis distance indicates the separation of the respective groups. It shows whether the model possesses an appropriate discriminatory power for differentiating between the two respective groups. The classification of cases was performed by means of the posterior classification probability, which is the probability that the respective case belongs to a particular group, i. e., mutants with near wild-type stability (H) or mutants with reduced stability (P). In developing this classification function the values of 1 and -1 were assigned to H and P mutants (see Table 9). The quality of the ADL-model was also determined by examining the percentage of good classification and the proportion between the cases and variables in the equation.

A simple linear and other more complex nonlinear model was obtaining using LMR and PLR as statistic techniques, respectively. The quality of the models was determined examining the statistic parameters of multivariable comparison of regression and cross-validation procedures. In this sense, the quality of models was determined by examining the regression coefficients (R), determination coefficients (R^2), Fisher-ratio's *p*-level [*p*(F)], standard deviations of the regression (s) and the leave-*one*-out (LOO) press statistics (q^2, s_{cv}).[82] In recent years, the LOO press statistics (e.g., q^2) have been used as a means of indicating predictive ability. Many authors consider high q^2 values (for instance, $q^2 > 0.5$) as indicator or even as the ultimate proof of the high-predictive power of a QSAR model. In a recent paper, Golbraikh and Tropsha demonstrated that a high value of LOO q^2 appears to be a necessary but not the sufficient condition for the model to have a high predictive power.[83]

In addition, to assess the robustness and predictive power of the found models, external prediction (test) sets were also used. This type of model validation is very important, if we take into consideration that the predictive ability of a QSAR model can

only be estimated using an external test set of compounds that was not used for building the model.[82, 83]

4. RESULTS AND DISCUSSION

4.1. Development of the Discriminant Function for Arc A-Mutants Classification.

The development of a discriminant function that permits the classification of mutants as near wild-type stability or reduced stability is a key of the present approach to describe the protein stability effects of a complete set of alanine substitutions in Arc repressor.

Here we considered a general data set of 53 A-mutants, 28 of them having near wild-type stability (1-28) and the rest being mutants with reduced stability (29-53). This data set was randomly divided into two subsets, one containing 41 mutants (21 having near wild-type stability and 20 of reduced stability) was used as a training set, and the other containing 12 mutants (7 having near wild-type stability and 5 of reduced stability) was used as a test set.

The principle of parsimony (Occam's razor) was taken into account as strategy for model selection. It were obtained non-stochastic and stochastic classification models (equations 21 and 22, respectively), each one was developed from protein structural depicting by means non-stochastic and stochastic bilinear indices. These are given below together with the statistical parameters of LDA:

$$\textbf{\textit{Class}} = -45.33 - 5.00\text{x}10^{-3}\,{}^{\mathbf{Z_1\text{-}ISA}}\boldsymbol{b}_0(\bar{x}_m, \bar{y}_m) - 1.00\text{x}10^{-3}\,{}^{\mathbf{Z_2\text{-}Z_3}}\boldsymbol{b}_6(\bar{x}_m, \bar{y}_m)$$

$$+2.00\,\text{x}10^{-3}\,{}^{\mathbf{Z_2\text{-}HPI}}\boldsymbol{b}_5(\bar{x}_m, \bar{y}_m) - 0.44\,{}^{\mathbf{ECI\text{-}HPI}}\boldsymbol{b}_2(\bar{x}_m, \bar{y}_m) \qquad (\mathbf{21})$$

$$N = 41 \quad \lambda = 0.24 \quad D^2 = 11.88 \quad F = 28.08 \quad p(F) < 0.0001$$

$$\textbf{\textit{Class}} = 24.80 - 5.00\text{x}10^{-3}\,{}^{\mathbf{Z_1\text{-}ISA}}\,{}^{s}\boldsymbol{b}_2(\bar{x}_m, \bar{y}_m) - 53.07\,{}^{\mathbf{ECI\text{-}HPI}}\,{}^{s}\boldsymbol{b}_0(\bar{x}_m, \bar{y}_m)$$

$$-0.47\,{}^{\mathbf{Z_2\text{-}ECI}}\,{}^{s}\boldsymbol{b}_1(\bar{x}_m, \bar{y}_m) - 0.15\,{}^{\mathbf{Z_2\text{-}HPI}}\,{}^{s}\boldsymbol{b}_6(\bar{x}_m, \bar{y}_m) \qquad (\mathbf{22})$$

N = 41 λ = 0.29 D^2 = 9.14 F = 21.61 p(F) < 0.0001

where λ is the Wilks's statistic, D^2 is the squared Mahalanobis distance and F is the Fisher ratio. The Mahalanobis distance indicates the separation of the respective groups. It shows whether the model possesses an appropriate discriminatory power for differentiating between the two respective groups.

These statistics indicate that models (Eqs. **21** and **22**) are appropriate for the discrimination of near wild-type stability/reduced stability mutants studied here. The non-stochastic-based QSAR obtained model has a positive predictive value of 100% (21/21) of near wild-type stability mutants and a negative predictive value of 100% (20/20) of reduced stability mutants in the training set, for an accuracy (global good classification) of 100% (41/41), while stochastic obtained model has a positive predictive value of 100% (21/21) of near wild-type stability mutants and a negative predictive value of 95.00% (19/20) of reduced stability mutants in the training set, for an accuracy (global good classification) of 97.56% (40/41).

Non-stochastic and stochastic models showed a high Matthew's correlation coefficients (MCC) of 1.00 and 0.95, respectively; MCC quantified the strength of the linear relation between the macromolecular descriptors and the classifications.[84] In Tables 9 and 10 we give the classification of mutants in the training set together with their posterior probabilities calculated from the Mahalanobis distance.

Table 9 and 10 comes about here (see end of the document)

The most important criterion to accept or not a discriminant models, such as models (Eqs **21** and **22**), is based on the statistics for the test set. Both models classify correctly 11 of 12 mutants, for an accuracy of 91.67%, with a MCC of 0.837. In Tables 9 and 10, we give the classification of mutants in the validation group. If we considered the data set and the test set (*full* set) the accuracy was 98.11% (52/53) and 96.23%

(51/53) for equation **21** and **22**, respectively, by using non-stochastic and stochastic bilinear indices in that order.

4.2. Development of the Regression Models for Melting Point Description.

The second step in modeling the stability effects of a complete set of A-substitution mutants was to find a way to predict the melting temperature (t_m) of such A-mutants of Arc repressor. With this aim, we compiled a data set of 48 proteins. Five A-mutants (49-53: VA22-st11, EA36-st11, IA37-st11, VA41-st11 and FA45-st11) were extracted due to their non-accurate t_m values (< 20 ^{o}C), which is not useful for MLR analysis.

By using the total non-stochastic and stochastic protein's bilinear indices and MLR analysis we developed the following QSSR [quantitative structure-stability relationship] lineal models to describe t_m for these A-mutants of the Arc repressor:

$$\begin{aligned} \boldsymbol{t_m}\ (^{o}\mathrm{C}) = {} & 51.07\ (\pm 0.58) - 8.58\ (\pm 1.43)\ {}^{\mathbf{Z_1\text{-}ISA}}\boldsymbol{b}_0(\bar{x}_m, \bar{y}_m) - 4.68\ (\pm 0.75)\ {}^{\mathbf{Z_2\text{-}Z_3}}\boldsymbol{b}_4(\bar{x}_m, \bar{y}_m) \\ & + 4.63\ (\pm 1.02)\ {}^{\mathbf{ISA\text{-}ECI}}\boldsymbol{b}_0(\bar{x}_m, \bar{y}_m) - 6.28\ (\pm 1.21)\ {}^{\mathbf{Z_1\text{-}Z_2}}\boldsymbol{b}_1(\bar{x}_m, \bar{y}_m) \\ & - 11.15\ (\pm 2.29)\ {}^{\mathbf{Z_1\text{-}HPI}}\boldsymbol{b}_1(\bar{x}_m, \bar{y}_m) - 7.77\ (\pm 2.80)\ {}^{\mathbf{Z_1\text{-}ECI}}\boldsymbol{b}_1(\bar{x}_m, \bar{y}_m) \qquad (\mathbf{23}) \end{aligned}$$

$N = 46\ R = 0.91\ R^2 = 0.83\ s = 3.9293\ q^2 = 0.73\ s_{cv} = 4.50\ F(6,39) = 33.08\ p<0.0001$

$$\begin{aligned} \boldsymbol{t_m}\ (^{o}\mathrm{C}) = {} & 51.02(\pm 0.63) - 5.29\ (\pm 1.25)\ {}^{\mathbf{Z1\text{-}ISA}\ s}\boldsymbol{b}_4(\bar{x}_m, \bar{y}_m) - 3.98(\pm 0.75)\ {}^{\mathbf{Z2\text{-}Z3}\ s}\boldsymbol{b}_4(\bar{x}_m, \bar{y}_m) \\ & + 9.08(\pm 1.73)\ {}^{\mathbf{Z2\text{-}HPI}\ s}\boldsymbol{b}_1(\bar{x}_m, \bar{y}_m) - 9.57(\pm 2.02)\ {}^{\mathbf{Z2\text{-}HPI}\ s}\boldsymbol{b}_3(\bar{x}_m, \bar{y}_m) \\ & - 2.17(\pm 0.68)\ {}^{\mathbf{ECI\text{-}HPI}\ s}\boldsymbol{b}_1(\bar{x}_m, \bar{y}_m) - 4.02(\pm 1.44)\ {}^{\mathbf{Z1\text{-}ISA}\ s}\boldsymbol{b}_1(\bar{x}_m, \bar{y}_m) \qquad (\mathbf{24}) \end{aligned}$$

$N = 46\ R = 0.90\ R^2 = 0.81\ s = 4.1941\ q^2 = 0.645\ s_{cv} = 5.01\ F(6,39) = 28.24\ p<0.0001$

where N is the size of the data set, R is the regression coefficient, s is the standard deviation of the regression, F is the Fischer ratio and q^2, s_{cv} are the squared correlation coefficient and the standard deviation of the cross-validation performed by the LOO procedure, respectively. Tables 7 and 8 give the observed and calculated t_m values from models developed by using non-stochastic and stochastic bilinear indices (Eqs. **23** and **24**, respectively) for the training set. Figures 1 and 2 are illustrated the linear relationships between observed and calculated t_m values by Eqs. **23** and **24**, respectively.

Equations **23** and **24** explain 83% and 81% of the variance of the experimental t_m, respectively. The predictive abilities of models are evidenced by the values of the LOO-press statistics ($q^2 > 0.5$ and s_{cv}).[82, 83] In developing these models only two mutants (1PA8-st6 and 45SA32-st11) were detected as statistical outliers.[85, 86] Outliers detection was carried out using the following standard statistical test: residual, standardized residual, studentized residual and Cook's distance.[86] Mutant (PA8) is only significantly more stable than wild type. The t_m of this mutant protein is about 15^{o}C higher than that of the wild-type parent (see Table 7), and the free energy of unfolding is increased by 2.9 kcal mol^{-1} compared with wild type.[34] In addition, different protein folding may be the reason for the lack of linear regression between protein's bilinear indices and stability (t_m) for these mutants; leading to a nonlinear dependence between t_m and protein's bilinear indices. In this case, other terms should be taken into consideration such as cooperative salt-bridges and hydrogen-bond formation, hydrophobic forces, steric terms, and so on. In this sense, far from strong quantitative correlations between stability and structural factors have been obtained in a previous study.[34] For example, when the set of t_m values were tested for linear correlations with fractional side-chain solvent accessibility, with changes in buried surface area, with average side-chain B-factors, and with the number of side-chain atoms or total

atoms within 6 Å of the atoms deleted by the alanine substitution, the pairwise correlation coefficient (r^2) ranges from 0.21 to 0.38.[34] Thus, even though most substitutions of alanine for hydrophobic-core residues are destabilizing, there is no simple relationship between the size of the replaced core residue and the destabilizing effect.[34]

Therefore, the use of other nonlinear models was required; a nonlinear model that retains linearity in the equation, but uses nonlinear methods to fit them. This is the piece-wise method,[80] which produces two linear equations by clustering observations into two groups according to their absolute magnitude. The best fitted non-stochastic (equations **25** and **26**) and stochastic (equations **27** and **28**) piecewise models were:

$$\boldsymbol{t_m}\,(^{o}\mathrm{C})_{<\mathrm{BKPT}} = 47.99 - 4.99^{\mathbf{Z1\text{-}Z2}}\boldsymbol{b}_1(\bar{x}_m,\bar{y}_m) - 9.07^{\mathbf{Z1\text{-}HPI}}\boldsymbol{b}_1(\bar{x}_m,\bar{y}_m) - 6.88^{\mathbf{Z1\text{-}ECI}}\boldsymbol{b}_1(\bar{x}_m,\bar{y}_m)$$
$$- 6.54^{\mathbf{Z1\text{-}ISA}}\boldsymbol{b}_0(\bar{x}_m,\bar{y}_m) - 4.25^{\mathbf{Z2\text{-}Z3}}\boldsymbol{b}_4(\bar{x}_m,\bar{y}_m) + 4.73^{\mathbf{ISA\text{-}ECI}}\boldsymbol{b}_0(\bar{x}_m,\bar{y}_m) \quad (\mathbf{25})$$

$$\boldsymbol{t_m}\,(^{o}\mathrm{C})_{>\mathrm{BKPT}} = 57.58 - 0.78^{\mathbf{Z1\text{-}Z2}}\boldsymbol{b}_1(\bar{x}_m,\bar{y}_m) - 1.47^{\mathbf{Z1\text{-}HPI}}\boldsymbol{b}_1(\bar{x}_m,\bar{y}_m) - 0.72^{\mathbf{Z1\text{-}ECI}}\boldsymbol{b}_1(\bar{x}_m,\bar{y}_m)$$
$$- 1.50^{\mathbf{Z1\text{-}ISA}}\boldsymbol{b}_0(\bar{x}_m,\bar{y}_m) - 1.29^{\mathbf{Z2\text{-}Z3}}\boldsymbol{b}_4(\bar{x}_m,\bar{y}_m) + 0.93^{\mathbf{ISA\text{-}ECI}}\boldsymbol{b}_0(\bar{x}_m,\bar{y}_m) \quad (\mathbf{26})$$

$N = 46 \quad R = 0.95 \quad R^2 = 0.91 \quad Bkpt = 51.86 \quad p < 0.0001$

$$\boldsymbol{t_m}\,(^{o}\mathrm{C})_{<\mathrm{BKPT}} = 47.49 - 4.93^{\mathbf{Z_1\text{-}Z_2}\,s}\boldsymbol{b}_1(\bar{x}_m,\bar{y}_m) - 3.42^{\mathbf{Z_1\text{-}HPI}\,s}\boldsymbol{b}_1(\bar{x}_m,\bar{y}_m)$$
$$- 2.40^{\mathbf{Z_1\text{-}ECI}\,s}\boldsymbol{b}_1(\bar{x}_m,\bar{y}_m) + 8{,}17^{\mathbf{Z_1\text{-}ISA}\,s}\boldsymbol{b}_0(\bar{x}_m,\bar{y}_m) - 10{,}07^{\mathbf{Z_2\text{-}Z_3}\,s}\boldsymbol{b}_4(\bar{x}_m,\bar{y}_m)$$
$$- 1{,}232^{\mathbf{ISA\text{-}ECI}\,s}\boldsymbol{b}_0(\bar{x}_m,\bar{y}_m) \quad (\mathbf{27})$$

$$\boldsymbol{t_m}\,(^{o}\mathrm{C})_{>\mathrm{BKPT}} = 58.28 - 1.13^{\mathbf{Z_1\text{-}Z_2}\,s}\boldsymbol{b}_1(\bar{x}_m,\bar{y}_m) + 0.31^{\mathbf{Z_1\text{-}HPI}\,s}\boldsymbol{b}_1(\bar{x}_m,\bar{y}_m)$$
$$- 0.84^{\mathbf{Z_1\text{-}ECI}\,s}\boldsymbol{b}_1(\bar{x}_m,\bar{y}_m) - 0.64^{\mathbf{Z_1\text{-}ISA}\,s}\boldsymbol{b}_0(\bar{x}_m,\bar{y}_m) + 0.32^{\mathbf{Z_2\text{-}Z_3}\,s}\boldsymbol{b}_4(\bar{x}_m,\bar{y}_m)$$
$$- 0.17^{\mathbf{ISA\text{-}ECI}\,s}\boldsymbol{b}_0(\bar{x}_m,\bar{y}_m) \quad (\mathbf{28})$$

$N = 46 \quad R = 0.96 \quad R^2 = 0.92 \quad Bkpt = 51.86 \quad p < 0.0001$

where R (piecewise regression coefficient), for gradual variance explanation, takes values ranging from 0 (non-piecewise regression) to 1 (explanation of 100% of variance). The probability of error after acceptance of the piece-wise hypothesis, p was checked for an absolute value > 0.05. The parameter break-point (Bkpt) is the t_m value, which mark the frontier between the two groups. The resultant regression coefficient suggested a highly significant piecewise linear correlation between observed and predicted values ($p < 0.05$). In Tables 11 and 12, we depict the observed, non-stochastic and stochastic calculated and residual values of t_m for the data set.

Table 11 and 12 comes about here (see end of the document)

The main difficulty of the linear piecewise regression is its limitation to predict new mutants whose stability profiles are unknown. The problem here is: which equation should be applied to a new mutant not considered in this study? The Bkpt value (51.86), perfectly agrees with an experimental scale previously proposed.[34] The same scale was used for grouping mutants into the two studied groups in our LDA approach. For this reason, we can use the LDA and piecewise models in combination to classify and to predict the stability of the mutant's Arc homodimers.

4.3. Structural Interpretation and Implication of Understanding Arc Folding

At present it is known that the folding of Arc repressor is influenced by different kinds of interactions.[34, 35] An overwhelming role is played by the van der waals forces. The hydrophobic interaction is another factor influencing the stability due to the hydrophobic nature of the Arc wild-type core. Another factor is related to electrostatic force, mainly due to intra and intersubunit salt bridges and hydrogen bonds.[34, 35] However, most of these factors are interrelated to each other, and it is difficult to determine the contribution of each one by separate. For instance, hydrophobic interaction is intimately related to van der Waals forces, and the electrostatic

interactions are also related to dispersion interactions, which are part of the van der Waals forces. In addition, Arc wild-type and its mutants showed a cooperative behaviour in folding/dimerization processes.

As can be observed in the obtained models, the included variables are related with the factors that influence on the stability and this one with the structural features of Arc dimer. In this sense, the protein's bilinear indices calculated using z_1-HPI, z_1-ISA, z_2-z_3, z_2-HPI, z_2-ECI couple values, as amino-acid (side-chain) properties-pairs are included in most of the developed models (see equations **21**-**28**). This pattern also displays when classification models are built using only one pair amino-acid (side-chain) properties and it is compared its global good classification (see Figure 3). These results draw individual significance of each one side-chain properties combination to explain variance of stability of A-mutants set.

Figure 3 comes about here (see end of the document)

These values are related to hydrophilicity (ISA, z_1), bulk-steric (z_2), and electronic (HPI, ECI and z_3) amino-acid side-chain properties (see Figure 3 and molecular descriptors include in equations **21**-**28**). For this reason, it is possible to determine the nature of the driving forces of the Arc repressor folding, e.g., hydrophobic, steric, or electronic. However, the preponderance of hydrophobic and electronic effects in the obtained equations (**21-28**) over other types of protein's bilinear indices clearly indicates the importance of the hydrophobic and electronic side-chain factor in the folding of Arc dimer. In fact, when we develop the final models (in this case Eqs. **21** and **22** with Q(%) of 100 and 97.56, respectively) by using at the same time whole set of bio-macromolecular descriptors (calculated with all weighting schedule), the result are better than when we used only one amino-acid side-chain property (best results archived with z_1-HPI and z_1-ISA based bilinear indices, which showed only 88% of

accuracy) to weighting every amino-acid in the Arc dimmer. This results evidence that Arc folding is a rather complicate process that depends to diverse process and the combinations of parameters (bilinear indices calculated with every pairs of amino-acid properties) are necessary to describe adequately the tm of these mutants' proteins (Eqs. **21** and **22**).

On the other hand, we plots Q(%) with specific orders of bilinear indices (Figure 4) in order to study the impact of vicinity in folding. The results show that orders in the range 0-13 are sufficient to explain the variance in tm and indices of high orders (k>13) are colineal. In the range 0-13, k = 1, 2 and 4 (Q = 95%) are the best of all orders as well as 6, 3 and 5 are the second best orders (Q = 93%, 88% and 88%, respectively). These results are very similar with the orders of bilinear indices that are in the equations **21** and **22**. In general form. It must be pointed out that developed equations (**21**-**28**) involve short-reaching ($k \leq 3$) and middle-reaching ($3 < k \leq 7$) protein's bilinear indices. Far-reaching (k = 8 or greater) bilinear indices were not considered like important to describe t_m, in complete agreement with the results obtained in the Figure 4. this is a logical results, because it has been well established that the inter-residue interaction (short, medium and long-range) play an important role in folding and stability of globular proteins.[11, 12] In fact, residues in ±1-6 vicinity (in the same- or in different-chain (amino-acid backbone) in Arc repressor dimmer) are the most relevant to describe the mutations of Arc native. This situation means that the change the amino-acid in the residue backbone do a more marked structural effects (inter-residue contacts) in the ±1-6-vicinity of the amino-acids. This situation means that the stability profile of wild-type Arc and its A-mutants results in topologic/topographic-controlled protein's backbone interactions.

Figure 4 comes about here (see end of the document)

4.4 Comparison with Other Computational Approaches.

Recently, some *in silico* techniques have been used to develop classification models that permits us compute biological stability for each A-mutant of Arc repressor.[29, 30, 44, 47, 87]

The relative comparison will be based on the kind of method use for deriving the QSAR and their statistical parameter, the explored molecular descriptors, the overall accuracy (%), Matthew's correlation coefficient and the validation method used. Table 13 describes the comparison between non-stochastic and stochastic macromolecular bilinear indices methods and others reported approaches for the stability prediction of A-mutants of Arc repressor.[29, 30, 44, 47, 87]

Table 13 comes about here (see end of document)

As can be seen, the accuracy in the training set (100% and 97.56%) of non-stochastic and stochastic bilinear indices based models were higher than of other reported LDA equations (for more details see Table 13). In addition, the Wilks' λ statistic for ours models was better than those reported in the others models.[29, 30, 44, 47, 87]

Validation of the models is the other major bottleneck in QSAR.[82, 83] One of the most popular validation criteria is internal cross-validation (leave-one-out, leave-n-out, leave-25%-out and so on). Nevertheless, there can exist a lack of correlation between the good results in internal cross-validation and the high predictive ability of QSAR models.[82, 83] Thus, the good high behavior in internal cross-validation appears to be the necessary but not the sufficient condition for the models to have a high predictive power. In this sense, Golbraikh and Tropsha emphasize that the predictive ability of a QSAR model can only be estimated using an external test set (external validation) of compounds that was not used for building the model and formulated a set

of criteria for evaluation of predictive ability of QSAR model.[82] In this case ours models show an accuracy of 91.67% for the test set. It is reasonable to expect some decrease in overall predictability of predicting sets with respect to training series for a simple reason; the model is developed to fit the points in training series, and therefore data points in predicting series are never used to develop it.

On the other hand, explained variance and LOO press statistics of non-stochastic bilinear indices linear model was higher than other ***TOMOCOMD-CARDD*** LMR equations reported by our group (see Table 14).[44, 47]

Table 14 comes about here (see end of document)

5. CONCLUDING REMARK

In this study a new set of bio-macromolecular descriptors relevant to protein QSAR/QSPR studies is present. These amino-acid based biochemical descriptors are based on the computation of bilinear maps on $\Re^n [b_{mk}(\bar{x}_m, \bar{y}_m) : \Re^n x \Re^n \rightarrow \Re]$ in canonical basis. Protein's bilinear indices are calculated from k^{th} power of non-stochastic and stochastic graph–theoretic electronic-contact matrices, M_m^k and ${}^sM_m^k$, respectively. Biochemical information is codified by using different pair combinations of amino-acid properties as weightings (*z*-values, side-chain isotropic surface area (ISA), amino-acids atomic charges (ECI) and hydrophathy index (Kyte-Doolittle scale; HPI). Their derivation is straightforward, and it is easy to interpret the QSARs/QSPRs that include them. We have shown here that the use of the protein's total bilinear indices is able to account for thermodynamic parameters for wild-type and mutant Arc proteins. The resulting quantitative models are significant from a statistical point of view. A LOO cross-validation procedure revealed that the QSA(S)R models had a good predictability. Protein's bilinear indices models compared favorably with several bio-macromolecular descriptors.

The models found to describe the stability profile of wild type Arc and its A-mutants include protein's bilinear indices accounting for hydrophobic (ISA, z_1), bulk-steric (z_2), and electronic (HPI, ECI and z_3) features of the studied molecules. These models using such combination of molecular descriptors are better than any other model that can be found by using only one type of the studied descriptors. We interpret these results as suggesting that many of the Arc mutations affect stability in more than one way and: by disrupting specific electronic interaction, by changing hydrophobic burial, and/or by changing the structure of the native or the denatured protein. Thus, we have proved that the combined use of z_1-HPI, z_1-ISA, z_2- z_3, z_2-HPI, z_2-ECI-protein's bilinear indices is an appropriated approach to QSSR studies. These models are not only good enough to predict thermodynamic parameter of the folding of mutants of Arc dimer repressor, but also permit the interpretation of the driving forces of such folding processes. The approach described here represents a novel and very promising way to bioinformatics research. We would expect computational protein science to have a similar effect on the search for new vaccines, receptors, drugs, and so on as molecular modelling and QSAR have had on search for new drugs.

Acknowledgement: Sadiel Ortega-Broche (O-B. S) acknowledges to Bioinformatics Research Center of Central University 'Marta Abreu'of Las Villas for kind hospitality during the 2006-2007. Yovani Marrero-Ponce (M-P. Y) acknowledges the Valencia University for kind hospitality during the first semester of 2008. M-P. Y thanks are given to the Valencia University, (Spain) for partial financial support as well as the program 'Estades Temporals per an Investigadors Convidats' for a fellowship to work at Pharmacy Faculty (2008). M-P. Y also thanks support from Spanish MEC (Project Reference: SAF2006-04698). Finally, but not less, CAMD-BIR Unit thanks are given to the research project called: "*Strengthening Postgraduate Education and Research in*

Pharmaceutical Sciences". This project is funded by the Flemish Interuniversity Council (VLIR) of Belgium.

6. REFERENCES

[1] EUFEPSAnnouncement., Conference on Optimizing Biotech Medicines: Rational Development of Therapeutic Proteins, Eur. J. Pharm. Sci. 15 (2002) 101-102.

[2] J. Saven, Combinatorial Protein Design, Curr. Opin. Struct. Biol. 12 (2002) 453-458.

[3] J. Mendes, R. Guerois, L. Serrano, Energy Estimation in Protein Design, Curr. Opin. Struct. Biol. 12 (2002) 441-446.

[4] D.N. Bolon, J.S. Marcus, S.A. Ross, S.L. Mayo, Prudent Modeling of Core Polar Residues in Computational Protein Design, J. Mol. Biol. 329 (2003) 611-622.

[5] L.L. Looger, M.A. Dwyer, J.J. Smith, H.W. Helling, Computational Design of Receptor and Sensor Proteins with Novel Functions, Nature 423 (2003) 185-190.

[6] L.X. Dang, K.M. Merz, P.A. Kollman, Free-energy calculations on protein stability: Thr-1573Val-157 mutation of T4 lysozyme, J. Am. Chem. Soc. 111 (1989) 8505-8508.

[7] M. Fernández, J. Caballero, L. Fernández, J.I. Abreu, M. Garriga, Protein radial distribution function (P-RDF) and Bayesian-Regularized Genetic Neural Networks for modeling protein conformational stability: Chymotrypsin inhibitor 2 mutants., J. Mol. Graphics Modell. 26 (2007) 748-759.

[8] H. Zhou, Y. Zhou, Stability scale and atomic solvation parameters extracted from 1023 mutation experiment., Proteins 49 (2002) 483-492.

[9] V. Munoz, L. Serrano, Development of the multiple sequence approximation within the AGADIR model of alpha-helix formation: comparison with Zimm–Bragg and Lifson–Roig formalisms, Biopolymers 41 (1997) 495-509.

[10] R. Guerois, J.E. Nielsen, L. Serrano, Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations, J. Mol. Biol. 320 (2002) 369-387.

[11] M.M. Gromiha, M. Oobatake, H. Kono, H. Uedaira, A. Sarai, Relationship between amino acid properties and protein stability: buried mutations, J. Protein Chem. 18 (1999) 565-578.

[12] M.M. Gromiha, M. Oobatake, H. Kono, H. Uedaira, A. Sarai, Role of structural and sequence information in the prediction of protein stability changes: comparison between buried and partially buried mutations, Protein Eng. 12 (1999) 549-555.

[13] C.M. Frenz, Neural network-based prediction of mutation-induced protein stability changes in staphylococcal nuclease at 20 residue positions, Proteins 59 (2005) 147-151.

[14] E. Capriotti, P. Fariselli, R. Casadio, A neural-network-based method for predicting protein stability changes upon single mutations, Bioinformatics 20 (2004) 63-68.

[15] E. Capriotti, P. Fariselli, R. Calabrese, R. Casadio, Prediction of protein stability changes from sequences using support vector machines, Bioinformatics 21 (2005) 54-58.
[16] R.C. Todeschini, V. (Ed.), Handbook of molecular descriptors., Germany, 2000.
[17] P.J. Flory, Principles of Polymer Chemistry, Cornell University Press, Itaha, 1953.
[18] A. Roy, C. Raychaudhury, A. Nandy, J. Biosci. 23 (1998) 55.
[19] J. Casanovas, J. Miro-Julia, F. Rossello, J. Math. Biol. 47 (2003) 1.
[20] P.M. Leong, S. Mogenthaler, Comput. Appl. Biosci. 12 (1995) 503.
[21] G.A. Arteca, J. Chem. Inf. Comput. Sci. 39 (1999) 550.
[22] M. Randic, A.T. Balaban, J. Chem. Inf. Comput. Sci. 43 (2003) 532.
[23] S. Hua, Z. Sun, Bioinformatics 17 (2001) 721.
[24] E. Estrada, A Protein Folding Degree Measure and Its Dependence on Crystal Packing, Protein Size, Secondary Structure, and Domain Structural Class, J. Chem. Inf. Comput. Sci. 44 (2004) 1238-1250.
[25] J. Kyte, R.F. Doolitle, A Simple Method for Displaying the Hydropathic Character of a Protein. , J. Mol. Biol. 157 (1982) 105-132.
[26] H. Gonzalez-Diaz, L.A. Torres-Gomez, Y. Guevara, M.S. Almeida, R. Molina, N. Castanedo, L. Santana, E. Uriarte, Markovian chemicals "in silico" design (MARCH-INSIDE), a promising approach for computer-aided molecular design III: 2.5D indices for the discovery of antibacterials, J Mol Model (Online) 11 (2005) 116-123.
[27] H. Gonzalez-Diaz, D. Vina, L. Santana, E. de Clercq, E. Uriarte, Stochastic entropy QSAR for the in silico discovery of anticancer compounds: Prediction, synthesis, and in vitro assay of new purine carbanucleosides, Bioorg Med Chem (2005).
[28] H. Gonzalez-Diaz, E. Uriarte, R. Ramos de Armas, Predicting stability of Arc repressor mutants with protein stochastic moments, Bioorg Med Chem 13 (2005) 323-331.
[29] H. Gonzalez-Diaz, E. Uriarte, Proteins QSAR with Markov average electrostatic potentials, Bioorg. Med. Chem. Lett. 15 (2005) 5088-5094.
[30] R.G.D. Ramos de Armas, H; Molina,R; and Uriarte E, Markovian Backbone Negentropies: Molecular Descriptors for Protein Research. I. Predicting Protein Stability in Arc Repressor Mutants, PROTEINS: Structure, Function, and Bioinformatics 56 (2004) 715-723.
[31] J.U. Bowie, R.T. Sauer, Equilibrium Dissociation and Unfolding of the Arc Repressor Dimmer, Biochemistry 28 (1989) 7139-7143.
[32] K.L. Knight, J.U. Bowie, A.K. Vershon, R.D. Kelley, R.T. Sauer, The Arc and Mnt Repressors: a New Class of Sequence Specific DNA-Binding Protein, J. Biol. Chem. 264 (1989) 3639-3642.
[33] M.E. Milla, M.B. Brown, R.T. Sauer, P22 Arc Repressor: Enhanced Expression of Unstable Mutants by Addition of Polar C-Terminal Sequences, Protein Sci. 2 (1993) 2198-2205.
[34] M.E. Milla, M.B. Brown, R.T. Sauer, Protein Stability Effects of a Complete Set of Alanine Substitutions in Arc Repressor, Struct. Biol. 1 (1994) 518-523.
[35] M.E. Milla, R.T. Saber, P22 Arc Repressor: Folding Kinetics of a Single Domain, Dimeric Protein. , Biochemistry 33 (1994) 1125-1133.
[36] A.K. Vershon, J.U. Bowie, T.M. Karplus, R.T. Sauer, Isolation and Analysis of Arc Repressor Mutants: Evidence for an Unusual Mechanism of DNA Binding, Proteins 1 (1986) 302-311.

[37] Y. Marrero-Ponce, Total and Local Quadratic Indices of the “Molecular Pseudograph`s Atom Adjacency Matrix”: Applications to the Prediction of Physical Properties of Organic Compounds., Molecules 8 (2003) 687-726.
[38] Y. Marrero-Ponce, Total and Local (Atom and Atom-Type) Molecular Quadratic Indices: Significance-Interpretation, Comparison to Other Molecular Descriptors and QSPR/QSAR Applications, Bioorg. Med. Chem. 12 (2004) 6351-6369.
[39] Y. Marrero-Ponce, J.A. Castillo-Garit, 3D-chiral Atom, Atom-type, and Total Non-stochastic and Stochastic Molecular Linear Indices and their Applications to Central Chirality Codification, J. Comput. Aided. Mol. Des. 19 (2005) 369-383.
[40] Y. Marrero-Ponce, H. González-Díaz, V. Romero-Zaldivar, F. Torrens, E.A. Castro, 3D-Chiral Quadratic Indices of the “Molecular Pseudograph’s Atom Adjacency Matrix” and their Application to Central Chirality Codification: Classification of ACE Inhibitors and Prediction of sigma-Receptor Antagonist Activities, Bioorg. Med. Chem. 12 (2004) 5331-5342.
[41] Y. Marrero-Ponce, J.A. Castillo-Garit, E. Olazabal, H.S. Serrano, A. Morales, N. Castanedo, F. Ibarra-Velarde, A. Huesca-Guillen, A.M. Sanchez, F. Torrens, E.A. Castro, Atom, atom-type and total molecular linear indices as a promising approach for bioorganic and medicinal chemistry: theoretical and experimental assessment of a novel method for virtual screening and rational design of new lead anthelmintic, Bioorg. Med. Chem. 13 (2005) 1005-1020.
[42] G.M. Casañola-Martin, M.T.H. Khan, Y. Marrero-Ponce, A. Ather, S. Sultan, F. Torrens, R. Rotondo, TOMOCOMD-CARDD Descriptors-Based Virtual Screening of Tyrosinase Inhibitors. Evaluation of Different Classification Model Combinations Using Bond-Based Linear Indices, Bioorg. Med. Chem. 15 (2007) 1483-1503.
[43] Y. Marrero-Ponce, F. Torrens, Y. Alvarado, R. Rotondo, Bond-Based Global and Local (Bond, Group and Bond-Type) Quadratic Indices and Their Applications to Computer-Aided Molecular Design. 1. QSPR Studies of Diverse Sets of Organic Chemicals, J. Comput. Aided Mol. Des. 20 (2006) 685-701.
[44] Y. Marrero-Ponce, R. Medina, E.A. Castro, R. de Armas, H. González, V. Romero, F. Torrens, Protein Quadratic Indices of the "Macromolecular Pseudograph's α-Carbon Atom Adjacency Matrix". 1. Prediction of Arc Repressor Alanine-mutant's Stability, Molecules 9 (2004) 1124-1147.
[45] Y. Marrero-Ponce, D. Nodarse, H.D. González, R. Ramos de Armas, V. Romero-Zaldivar, F. Torrens, E. Castro, Nucleic Acid Quadratic Indices of the "Macromolecular Graph's Nucleotides Adjacency Matrix". Modeling of Footprints after the Interaction of Paromomycin with the HIV-1 Ψ-RNA Packaging Region, Int. J. Mol. Sci. 5 (2004) 276-293.
[46] Y. Marrero Ponce, J.A. Castillo Garit, D. Nodarse, Linear indices of the 'macromolecular graph's nucleotides adjacency matrix' as a promising approach for bioinformatics studies. Part 1: prediction of paromomycin's affinity constant with HIV-1 psi-RNA packaging region, Bioorg. Med. Chem. 13 (2005) 3397-3404.
[47] Y. Marrero-Ponce, R. Medina-Marrero, J.A. Castillo-Garit, V. Romero-Zaldivar, F. Torrens, E.A. Castro, Protein linear indices of the 'macromolecular pseudograph alpha-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 (2005) 3003-3015.

[48] Y. Marrero-Ponce, F. Torrens, R. García-Domenech, S.E. Ortega-Broche, V. Romero-Zaldivar, Novel 2D TOMOCOMD-CARDD Descriptors: Atom-based Stochastic and non-Stochastic Bilinear Indices and their QSPR Applications, J. Math. Chem. DOI10.1007/s10910-008-9389-0 (2008).
[49] Y. Marrero-Ponce, M.T.H. Khan, G.M. Casañola-Martin, A. Ather, K.M. Khan, M.K. Khan., F. Torrens, R. Rotondo, Bond-Based 2D TOMOCOMD-CARDD Approach for Drug Discovery: Aiding Decision-Making in in silico Selection of New Lead Tyrosinase Inhibitors, J. Comput. Aided Mol. Des. 21 (2007) 167-188.
[50] J.A. Castillo-Garit, Y. Marrero-Ponce, F. Torrens, R. Rotondo, Atom-based Stochastic and non-Stochastic 3D-Chiral Bilinear Indices and their Applications to Central Chirality Codification., J. Mol. Graph. Model. 26 (2007) 32-47.
[51] Y. Marrero-Ponce, M.T.H. Khan, G.M. Casañola-Martin, A. Ather, M.N. Sultankhodzhaev, F. Torrens, R. Rotondo, Prediction of Tyrosinase Inhibition Spectra for Chemicals Using Novel Atom-Based Bilinear Indices, ChemMedChem 2 (2007) 449-478.
[52] Y. Marrero-Ponce, A. Meneses-Marcel, Y. Machado-Tugores, J.A. Escario, A. Gómez-Barrio, D. Montero Pereira, J.J. Nogal-Ruiz, V.J. Arán, A.R. Martínez-Fernández, F. Torrens, R. Rotondo, Predicting Antitrichomonal Activity: A Computational Screening Using Atom-Based Bilinear Indices and Experimental Proofs, Bioorg. Med. Chem. 14 (2006) 6502-6524.
[53] D.J. Klein, Graph Theoretically Formulated Electronic–Structure Theory. , Internet Electron. J. Mol. Des. 2 (2003) 814-834.
[54] Y. Marrero-Ponce, M.A. Cabrera, V. Romero, D.H. González, F. Torrens, A New Topological Descriptors Based Model for Predicting Intestinal Epithelial Transport of Drugs in Caco-2 Cell Culture, J. Pharm. Pharm. Sci 7 (2004) 186-199.
[55] S. Hellberg, M. Sjöström, B. Skagerberg, S. Wold, Peptide Quantitative Structure-Activity Relationship, a Multivariate Approach, J. Med. Chem. 30 (1987) 1126-1135.
[56] E.R.D.I.W.J. Collantes, Amino Acid Side Chain Descriptors for Quantitative Structure-Activity Relationship Studies of Peptide Analogues, J. Med. Chem. 38 (1995,) 2705-2713.
[57] J. Kyte, R.F. Doolitle, A Simple Method for Displaying the Hydropathic Character of a Protein, J. Mol. Biol. 157 (1982) 105-132.
[58] T.P. Hoop, K.R. Woods, Prediction of Protein Antigenic Determinants from Amino Scid Sequences, Proc. Natl. Acad. Sci. U.S.A. 78 (1981) 3824-3828.
[59] D.H. Rouvray (Ed.), In Chemical Applications of Graph Theory, Academic Press, London, 1976.
[60] N. Trinajstić (Ed.), Chemical Graph Theory, 2nd ed.; 1992 ed., CRC Press, Boca Raton, FL, 1983.
[61] a.O.E.P. I. Gutman (Ed.), Mathematical Concepts in Organic Chemistry, Berlin, 1986.
[62] E. Estrada, G. Patlewicz, On the Usefulness of Graph-theoretic Descriptors in Predicting Theoretical Parameters. Phototoxicity of Polycyclic Aromatic Hydrocarbons (PAHs), Croat. Chem. Acta 77 (2004) 203-211.
[63] Y. Marrero-Ponce, Linear Indices of the "Molecular Pseudograph's Atom Adjacency Matrix": Definition, Significance-Interpretation and Application to QSAR Analysis of Flavone Derivatives as HIV-1 Integrase Inhibitors, J. Chem. Inf. Comput. Sci 44 (2004) 2010-2026.

[64] C.H. Edwards, D.E. Penney, Elementary Linear Algebra, Prentice-Hall, Englewood Cliffs, New Jersey, USA 1988.
[65] N. Jacobson (Ed.), Basic Algebra I 2nd Edition ed., Ed. W.H. Freeman and Company, New York, 1985.
[66] M.P.H. K. F. Riley, and S. J. Vence (Ed.), Mathematical Methods for Physics and Engineering. 1998.
[67] E. Hernández (Ed.), Álgebra y Geometría, Universidad Autonoma de Madrid, Madrid, Spain, 1987.
[68] J. de Burgos-Román (Ed.), Álgebra y Geometría Cartesíana, 2da Edición ed., McGraw-Hill Interamericana 2000.
[69] J. de Burgos-Román (Ed.), Curso de Álgebra y Geometría, Alambra Longman, Madrid, Spain, 1994.
[70] G. Werner (Ed.), Linear Algebra, 4th ed., Springer-Verlag, New York, USA, 1981.
[71] M. Randić, Generalized Molecular Descriptors, J. Math. Chem. 7 (1991) 155-168.
[72] P.G.M. D.Walker, J. Am. Chem. Soc. 115 (1993) 12423.
[73] Y.R. Marrero-Ponce, V. , TOMOCOMD software (TOpological MOlecular COMputer Design) for Windows, version 1.0, Central University of Las Villas, 2002.
[74] A.K.B. Vershon, J. U.; Karplus, T. M.; Sauer, R. T. , Isolation and Analysis of Arc Repressor Mutants: Evidence for an Unusual Mechanism of DNA Binding., Proteins. 1 (1986,) 302-311.
[75] J.U.S. Bowie, R. T. , Identifying Determinants of Folding and Activity for a Protein of Unknown Structure. , Proc. Natl. Acad. Sci. USA. 86 (1989) 2152-2156.
[76] T. Alber, Mutational Effects on Protein Stability, Annu. Rev. Biochem 58 (1989) 765-798.
[77] D.P. Goldenberg, Genetic Studies of Proteins Stability and Mechanisms of Folding, Annu. Rev. Biophys. Biophys. Chem 17 (1988),481-507.
[78] B.W. Matthews, Structural and Genetic Analysis of Protein Stability, Annu. Rev. Biochem. 62, (1993) 139-160.
[79] D. Shortle, Denature States of Proteins and Their Roles in Folding and Stability, Curr. Opin. Struct. Biol. 3 (1993) 66-74.
[80] STATISTICA, Statsoft, Inc, 1999.
[81] J.W. McFarland, D.J. Gans, Linear Discrminant Analysis and Cluster Significance Analysis, Pergamon Press, Oxford, 1990.
[82] S.E. Wold, L., Statistical Validation of QSAR Results. Validation Tools, In Chemometric Methods in Molecular Design, New York, 1995.
[83] A. Golbraikh, A. Tropsha, Beware of q2, J. Mol. Graphics. Mod. 20 (2002) 269-276.
[84] P. Baldi, S. Brunak, Y. Chauvin, H. Nielsen, Assessing the Accuracy of Prediction Algorithms for Classification: An Overview, Bioinformatics 16 (2000) 412-424.
[85] M.T.D. Cronin, T.W. Schultz, J. Mol. Struct. (Theochem.) 622 (2003) 39.
[86] D.A. Belsey, E. Kuh, R.E. Welsch (Eds.), Regression Diagnostics, New York, 1980.
[87] H. Gonzalez-Dıaz, E. Uriarte, R. Ramos de Armas, Predicting stability of Arc repressor mutants with protein stochastic moments, Bioorg. Med. Chem. 13 (2005) 323-331.

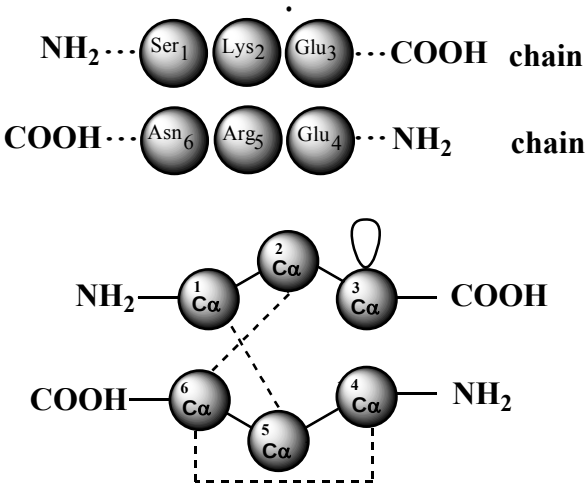
ANNEXES

(Tables, figures and schemes should be inserted into main text)

Table 1. Descriptors for the natural amino-acids.

Amino-acids		z-scale[55, 56]			Hydrophobicity Scale (HPI) (Kyte-Doolittle)[57]	ISA[56]	ECI[56]
		z_1	z_2	z_3			
Ala	**A**	0.07	-1.73	0.09	1.8	62.90	0.05
Val	**V**	-2.69	-2.53	-1.29	4.2	120.91	0.07
Leu	**L**	-4.19	-1.03	-0.98	3.8	154.35	0.01
Ile	**I**	-4.44	-1.68	-1.03	4.5	149.77	0.09
Pro	**P**	-1.22	0.88	2.23	-1.6	122.35	0.16
Phe	**F**	-4.92	1.30	0.45	2.8	189.42	0.14
Trp	**W**	-4.75	3.65	0.85	-0.9	179.16	1.08
Met	**M**	-2.49	-0.27	-0.41	1.9	132.22	0.34
Lys	**K**	2.84	1.41	-3.14	-3.9	102.78	0.53
Arg	**R**	2.88	2.52	-3.44	-4.5	52.98	1.69
His	**H**	2.41	1.74	1.11	-3.2	87.38	0.56
Gly	**G**	2.23	-5.36	0.30	-0.4	19.93	0.02
Ser	**S**	1.96	-1.63	0.57	-0.8	19.75	0.56
Thr	**T**	0.92	-2.09	-1.40	-0.7	59.44	0.65
Cys	**C**	0.71	-0.97	4.13	2.5	78.51	0.15
Tyr	**Y**	-1.39	2.32	0.01	-1.3	132.16	0.72
Asn	**N**	3.22	1.45	0.84	-3.5	17.87	1.31
Gln	**Q**	2.18	0.53	-1.14	-3.5	19.53	1.36
Asp	**D**	3.64	1.13	2.36	-3.5	18.46	1.25
Glu	**E**	3.08	0.39	-0.07	-3.5	30.19	1.31

Table 2. Representation of two interacting polypeptide chains and its associated pseudograph and macromolecular vector.



Macromolecular 'Pseudograph' (G_m) of the α-Carbon Atoms (Polypeptide's backbone)

Here we consider both covalent interaction (peptidic bond between amino-acid shown with solid line) and non-covalent interaction (salt bridge and hydrogen-bond shown with dashed line) between amino-acid side chains (R-groups) in a same polypeptidic chain or different. Loop in third position (Glu_3) means hydrogen-bond between amino-acid main chain and its side-chain.

Macromolecular vector:

$\bar{x}_m = [S \quad K \quad E \quad E \quad R \quad N] \in R^6$

In the definition of the $\bar{x}_m$, as macromolecular vector, the one letter symbol of the amino-acids indicates the corresponding side-chain amino-acid property, e.g., z_1-values. That is to say, if we write S it means $z_1(S)$, z_1-values or some amino-acid property, which characterizes each side chain in the polypeptide. Therefore, if we use the canonical bases of R^6, the coordinates of any vector $\bar{x}_m$ coincide with the components of that macromolecular vector.

$[X_m]^T = [S \quad K \quad E \quad E \quad R \quad N]$

$[X_m]^T$ = transposed of $[X_m]$ and it means the vector of the coordinates of $\bar{x}_m$ in the canonical basis of R^6 (an 1x6 matrix)

$[X_m]$: vector of coordinates of $\bar{x}_m$ in the canonical basis of R^6 (a 6x1matrix).

$\bar{x}_m$, $\bar{y}_m$ components are z_1 and z_2-values respectively.

$\bar{x}_m$=[1.96 2.84 3.08 3.08 2.88 3.22]

$\bar{y}_m$=[-1.63 1.41 0.39 0.39 2.52 1.45]

Table 3. The zero (k = 0), first (k = 1) and second (k = 2) powers of the total non-stochastic and stochastic graph–theoretic electronic-contact matrices of G_m, respectively.

Order (k)	Non-Stochastic	Stochastic
$k = 0$	$\begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$	$\begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$
$k = 1$	$\begin{bmatrix} 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 1 & 0 \end{bmatrix}$	$\begin{bmatrix} 0 & \frac{1}{2} & 0 & 0 & \frac{1}{2} & 0 \\ \frac{1}{3} & 0 & \frac{1}{3} & 0 & 0 & \frac{1}{3} \\ 0 & \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{3} & 0 & 0 & \frac{1}{3} & 0 & \frac{1}{3} \\ 0 & \frac{1}{3} & 0 & \frac{1}{3} & \frac{1}{3} & 0 \end{bmatrix}$
$k = 2$	$\begin{bmatrix} 2 & 0 & 1 & 1 & 0 & 2 \\ 0 & 3 & 1 & 1 & 2 & 0 \\ 1 & 1 & 2 & 0 & 0 & 1 \\ 1 & 1 & 0 & 2 & 1 & 1 \\ 0 & 2 & 0 & 1 & 3 & 1 \\ 2 & 0 & 1 & 1 & 1 & 3 \end{bmatrix}$	$\begin{bmatrix} \frac{1}{3} & 0 & \frac{1}{6} & \frac{1}{6} & 0 & \frac{1}{3} \\ 0 & \frac{3}{7} & \frac{1}{7} & \frac{1}{7} & \frac{2}{7} & 0 \\ \frac{1}{5} & \frac{1}{5} & \frac{2}{5} & 0 & 0 & \frac{1}{5} \\ \frac{1}{6} & \frac{1}{6} & 0 & \frac{1}{3} & \frac{1}{6} & \frac{1}{6} \\ 0 & \frac{2}{7} & 0 & \frac{1}{7} & \frac{3}{7} & \frac{1}{7} \\ \frac{1}{4} & 0 & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & \frac{3}{8} \end{bmatrix}$

Table 4. Values of non-stochastic and stochastic total bilinear indices for two interacting peptides (S K E E R N) used as example above (see also Table 2 and 3).

Non-Stochastic Total Bilinear Indices

$$b_{m0} = \sum_{i=1}^{n}\sum_{j=1}^{n} {}^{0}m_{ij}x_m^i y_m^j = [X_m]^T M_m^0 [Y_m] = [S \quad K \quad E \quad E \quad R \quad N] \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} S \\ K \\ E \\ E \\ R \\ N \end{bmatrix} = 15.1386$$

$$b_{m1} = \sum_{i=1}^{n}\sum_{j=1}^{n} {}^{1}m_{ij}x_m^i y_m^j = [X_m]^T M_m^1 [Y_m] = [S \quad K \quad E \quad E \quad R \quad N] \begin{bmatrix} 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 1 & 0 \end{bmatrix} \begin{bmatrix} S \\ K \\ E \\ E \\ R \\ N \end{bmatrix} = 40.586$$

$$b_{m2} = \sum_{i=1}^{n}\sum_{j=1}^{n} {}^{2}m_{ij}x_m^i y_m^j = [X_m]^T M_m^2 [Y_m] = [S \quad K \quad E \quad E \quad R \quad N] \begin{bmatrix} 2 & 0 & 1 & 1 & 0 & 2 \\ 0 & 3 & 1 & 1 & 2 & 0 \\ 1 & 1 & 2 & 0 & 0 & 1 \\ 1 & 1 & 0 & 2 & 1 & 1 \\ 0 & 2 & 0 & 1 & 3 & 1 \\ 2 & 0 & 1 & 1 & 1 & 3 \end{bmatrix} \begin{bmatrix} S \\ K \\ E \\ E \\ R \\ N \end{bmatrix} = 98.8378$$

Stochastic Total Bilinear Indices

$${}^{s}b_{m0} = \sum_{i=1}^{n}\sum_{j=1}^{n} {}^{0}sm_{ij}x_m^i y_m^j = [X_m]^T {}^{s}M_m^0 [Y_m] = [S \quad K \quad E \quad E \quad R \quad N] \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} S \\ K \\ E \\ E \\ R \\ N \end{bmatrix} = 15.1386$$

$${}^{s}b_{m1} = \sum_{i=1}^{n}\sum_{j=1}^{n} {}^{1}sm_{ij}x_m^i y_m^j = [X_m]^T {}^{s}M_m^1 [Y_m] = [S \quad K \quad E \quad E \quad R \quad N] \begin{bmatrix} 0 & \frac{1}{2} & 0 & 0 & \frac{1}{2} & 0 \\ \frac{1}{3} & 0 & \frac{1}{3} & 0 & 0 & \frac{1}{3} \\ 0 & \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{3} & 0 & 0 & \frac{1}{3} & 0 & \frac{1}{3} \\ 0 & \frac{1}{3} & 0 & \frac{1}{3} & \frac{1}{3} & 0 \end{bmatrix} \begin{bmatrix} S \\ K \\ E \\ E \\ R \\ N \end{bmatrix} = 17.7744$$

$${}^{s}b_{m2} = \sum_{i=1}^{n}\sum_{j=1}^{n} {}^{2}sm_{ij}x_m^i y_m^j = [X_m]^T {}^{s}M_m^2 [Y_m] = [S \quad K \quad E \quad E \quad R \quad N] \begin{bmatrix} \frac{1}{3} & 0 & \frac{1}{6} & \frac{1}{6} & 0 & \frac{1}{3} \\ 0 & \frac{3}{7} & \frac{1}{7} & \frac{1}{7} & \frac{2}{7} & 0 \\ \frac{1}{5} & \frac{1}{5} & \frac{2}{5} & 0 & 0 & \frac{1}{5} \\ \frac{1}{6} & \frac{1}{6} & 0 & \frac{1}{3} & \frac{1}{6} & \frac{1}{6} \\ 0 & \frac{2}{7} & 0 & \frac{1}{7} & \frac{3}{7} & \frac{1}{7} \\ \frac{1}{4} & 0 & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & \frac{3}{8} \end{bmatrix} \begin{bmatrix} S \\ K \\ E \\ E \\ R \\ N \end{bmatrix} = 14.5728207$$

Table 5. The zero (k = 0), first (k = 1) and second (k = 2) powers of the local non-stochastic and stochastic graph–theoretic electronic-contact matrices of G_m, respectively.

The zero, first and second powers of the local (amino-acid) non-stochastic matrices

$$M^0(G_m,S)=\begin{bmatrix}1&0&0&0&0&0\\0&0&0&0&0&0\\0&0&0&0&0&0\\0&0&0&0&0&0\\0&0&0&0&0&0\\0&0&0&0&0&0\end{bmatrix} \quad M^1(G_m,S)=\begin{bmatrix}0&\frac{1}{2}&0&0&\frac{1}{2}&0\\\frac{1}{2}&0&0&0&0&0\\0&0&0&0&0&0\\0&0&0&0&0&0\\\frac{1}{2}&0&0&0&0&0\\0&0&0&0&0&0\end{bmatrix} \quad M^2(G_m,S)=\begin{bmatrix}2&0&\frac{1}{2}&\frac{1}{2}&0&1\\0&0&0&0&0&0\\\frac{1}{2}&0&0&0&0&0\\\frac{1}{2}&0&0&0&0&0\\0&0&0&0&0&0\\1&0&0&0&0&0\end{bmatrix}$$

$$M^0(G_m,K)=\begin{bmatrix}0&0&0&0&0&0\\0&1&0&0&0&0\\0&0&0&0&0&0\\0&0&0&0&0&0\\0&0&0&0&0&0\\0&0&0&0&0&0\end{bmatrix} \quad M^1(G_m,K)=\begin{bmatrix}0&\frac{1}{2}&0&0&0&0\\\frac{1}{2}&0&\frac{1}{2}&0&0&\frac{1}{2}\\0&\frac{1}{2}&0&0&0&0\\0&0&0&0&0&0\\0&0&0&0&0&0\\0&\frac{1}{2}&0&0&0&0\end{bmatrix} \quad M^2(G_m,K)=\begin{bmatrix}0&0&0&0&0&0\\0&3&\frac{1}{2}&\frac{1}{2}&1&0\\0&\frac{1}{2}&0&0&0&0\\0&\frac{1}{2}&0&0&0&0\\0&1&0&0&0&0\\0&0&0&0&0&0\end{bmatrix}$$

$$M^0(G_m,E)=\begin{bmatrix}0&0&0&0&0&0\\0&0&0&0&0&0\\0&0&1&0&0&0\\0&0&0&0&0&0\\0&0&0&0&0&0\\0&0&0&0&0&0\end{bmatrix} \quad M^1(G_m,E)=\begin{bmatrix}0&0&0&0&0&0\\0&0&\frac{1}{2}&0&0&0\\0&\frac{1}{2}&1&0&0&0\\0&0&0&0&0&0\\0&0&0&0&0&0\\0&0&0&0&0&0\end{bmatrix} \quad M^2(G_m,E)=\begin{bmatrix}0&0&\frac{1}{2}&0&0&0\\0&0&\frac{1}{2}&0&0&0\\\frac{1}{2}&\frac{1}{2}&2&0&0&\frac{1}{2}\\0&0&0&0&0&0\\0&0&0&0&0&0\\0&0&\frac{1}{2}&0&0&0\end{bmatrix}$$

$$M^0(G_m,E)=\begin{bmatrix}0&0&0&0&0&0\\0&0&0&0&0&0\\0&0&0&0&0&0\\0&0&0&1&0&0\\0&0&0&0&0&0\\0&0&0&0&0&0\end{bmatrix} \quad M^1(G_m,E)=\begin{bmatrix}0&0&0&0&0&0\\0&0&0&0&0&0\\0&0&0&0&0&0\\0&0&0&0&\frac{1}{2}&\frac{1}{2}\\0&0&0&\frac{1}{2}&0&0\\0&0&0&\frac{1}{2}&0&0\end{bmatrix} \quad M^2(G_m,E)=\begin{bmatrix}0&0&0&\frac{1}{2}&0&0\\0&0&0&\frac{1}{2}&0&0\\0&0&0&0&0&0\\\frac{1}{2}&\frac{1}{2}&0&2&\frac{1}{2}&\frac{1}{2}\\0&0&0&\frac{1}{2}&0&0\\0&0&0&\frac{1}{2}&0&0\end{bmatrix}$$

$$M^0(G_m,R)=\begin{bmatrix}0&0&0&0&0&0\\0&0&0&0&0&0\\0&0&0&0&0&0\\0&0&0&0&0&0\\0&0&0&0&1&0\\0&0&0&0&0&0\end{bmatrix} \quad M^1(G_m,R)=\begin{bmatrix}0&0&0&0&\frac{1}{2}&0\\0&0&0&0&0&0\\0&0&0&0&0&0\\0&0&0&0&\frac{1}{2}&0\\\frac{1}{2}&0&0&\frac{1}{2}&0&\frac{1}{2}\\0&0&0&0&\frac{1}{2}&0\end{bmatrix} \quad M^2(G_m,R)=\begin{bmatrix}0&0&0&0&0&0\\0&0&0&0&1&0\\0&0&0&0&0&0\\0&0&0&0&\frac{1}{2}&0\\0&1&0&\frac{1}{2}&3&\frac{1}{2}\\0&0&0&0&\frac{1}{2}&0\end{bmatrix}$$

$$M^0(G_m,N)=\begin{bmatrix}0&0&0&0&0&0\\0&0&0&0&0&0\\0&0&0&0&0&0\\0&0&0&0&0&0\\0&0&0&0&0&0\\0&0&0&0&0&1\end{bmatrix} \quad M^1(G_m,N)=\begin{bmatrix}0&0&0&0&0&0\\0&0&0&0&0&\frac{1}{2}\\0&0&0&0&0&0\\0&0&0&0&0&\frac{1}{2}\\0&0&0&0&0&\frac{1}{2}\\0&\frac{1}{2}&0&\frac{1}{2}&\frac{1}{2}&0\end{bmatrix} \quad M^2(G_m,N)=\begin{bmatrix}0&0&0&0&0&1\\0&0&0&0&0&0\\0&0&0&0&0&\frac{1}{2}\\0&0&0&0&0&\frac{1}{2}\\0&0&0&0&0&\frac{1}{2}\\1&0&\frac{1}{2}&\frac{1}{2}&\frac{1}{2}&3\end{bmatrix}$$

The zero, first and second powers of the local (amino-acid) stochastic matrices

$$M^0(G_m,S)=\begin{bmatrix}1&0&0&0&0&0\\0&0&0&0&0&0\\0&0&0&0&0&0\\0&0&0&0&0&0\\0&0&0&0&0&0\\0&0&0&0&0&0\end{bmatrix} \quad M^1(G_m,S)=\begin{bmatrix}0&\frac{1}{4}&0&0&\frac{1}{4}&0\\\frac{1}{6}&0&0&0&0&0\\0&0&0&0&0&0\\0&0&0&0&0&0\\\frac{1}{6}&0&0&0&0&0\\0&0&0&0&0&0\end{bmatrix} \quad M^2(G_m,S)=\begin{bmatrix}\frac{1}{3}&0&\frac{1}{12}&\frac{1}{12}&0&\frac{1}{6}\\\frac{1}{6}&0&0&0&0&0\\\frac{1}{10}&0&0&0&0&0\\\frac{1}{12}&0&0&0&0&0\\\frac{1}{8}&0&0&0&0&0\\0&0&0&0&0&0\end{bmatrix}$$

$$M^0(G_m,K)=\begin{bmatrix}0&0&0&0&0&0\\0&1&0&0&0&0\\0&0&0&0&0&0\\0&0&0&0&0&0\\0&0&0&0&0&0\\0&0&0&0&0&0\end{bmatrix} \quad M^1(G_m,K)=\begin{bmatrix}0&\frac{1}{4}&0&0&0&0\\\frac{1}{6}&0&\frac{1}{6}&0&0&\frac{1}{6}\\0&\frac{1}{4}&0&0&0&0\\0&0&0&0&0&0\\0&0&0&0&0&0\\0&\frac{1}{6}&0&0&0&0\end{bmatrix} \quad M^2(G_m,K)=\begin{bmatrix}0&0&0&0&0&0\\0&\frac{3}{7}&\frac{1}{14}&\frac{1}{14}&\frac{1}{7}&0\\0&\frac{1}{10}&0&0&0&0\\0&\frac{1}{12}&0&0&0&0\\0&\frac{1}{7}&0&0&0&0\\0&0&0&0&0&0\end{bmatrix}$$

$$M^0(G_m,E)=\begin{bmatrix}0&0&0&0&0&0\\0&0&0&0&0&0\\0&0&1&0&0&0\\0&0&0&0&0&0\\0&0&0&0&0&0\\0&0&0&0&0&0\end{bmatrix} \quad M^1(G_m,E)=\begin{bmatrix}0&0&0&0&0&0\\0&0&\frac{1}{6}&0&0&0\\0&\frac{1}{4}&\frac{1}{2}&0&0&0\\0&0&0&0&0&0\\0&0&0&0&0&0\\0&0&0&0&0&0\end{bmatrix} \quad M^2(G_m,E)=\begin{bmatrix}0&0&\frac{1}{12}&0&0&0\\0&0&\frac{1}{14}&0&0&0\\\frac{1}{10}&\frac{1}{10}&\frac{2}{5}&0&0&\frac{1}{10}\\0&0&0&0&0&0\\0&0&0&0&0&0\\0&0&\frac{1}{16}&0&0&0\end{bmatrix}$$

Table 5. *Cont.*

$$M^0(G_m,E)=\begin{bmatrix}0&0&0&0&0&0\\0&0&0&0&0&0\\0&0&0&0&0&0\\0&0&0&1&0&0\\0&0&0&0&0&0\\0&0&0&0&0&0\end{bmatrix}$$

$$M^1(G_m,E)=\begin{bmatrix}0&0&0&0&0&0\\0&0&0&0&0&0\\0&0&0&0&0&0\\0&0&0&0&\frac{1}{4}&\frac{1}{4}\\0&0&0&\frac{1}{6}&0&0\\0&0&0&\frac{1}{6}&0&0\end{bmatrix}$$

$$M^2(G_m,E)=\begin{bmatrix}0&0&0&\frac{1}{12}&0&0\\0&0&0&\frac{1}{14}&0&0\\0&0&0&0&0&0\\\frac{1}{12}&\frac{1}{12}&0&\frac{1}{3}&\frac{1}{12}&\frac{1}{12}\\0&0&0&\frac{1}{14}&0&0\\0&0&0&\frac{1}{16}&0&0\end{bmatrix}$$

$$M^0(G_m,R)=\begin{bmatrix}0&0&0&0&0&0\\0&0&0&0&0&0\\0&0&0&0&0&0\\0&0&0&0&0&0\\0&0&0&0&1&0\\0&0&0&0&0&0\end{bmatrix}$$

$$M^1(G_m,R)=\begin{bmatrix}0&0&0&0&\frac{1}{4}&0\\0&0&0&0&0&0\\0&0&0&0&0&0\\0&0&0&0&\frac{1}{4}&0\\\frac{1}{6}&0&0&\frac{1}{6}&0&\frac{1}{6}\\0&0&0&0&\frac{1}{6}&0\end{bmatrix}$$

$$M^2(G_m,R)=\begin{bmatrix}0&0&0&0&0&0\\0&0&0&0&\frac{1}{7}&0\\0&0&0&0&0&0\\0&0&0&0&\frac{1}{12}&0\\0&\frac{1}{7}&0&\frac{1}{14}&\frac{3}{7}&\frac{1}{14}\\0&0&0&0&\frac{1}{16}&0\end{bmatrix}$$

$$M^0(G_m,N)=\begin{bmatrix}0&0&0&0&0&0\\0&0&0&0&0&0\\0&0&0&0&0&0\\0&0&0&0&0&0\\0&0&0&0&0&0\\0&0&0&0&0&1\end{bmatrix}$$

$$M^1(G_m,N)=\begin{bmatrix}0&0&0&0&0&0\\0&0&0&0&0&\frac{1}{6}\\0&0&0&0&0&0\\0&0&0&0&0&\frac{1}{4}\\0&0&0&0&0&\frac{1}{6}\\0&\frac{1}{6}&0&\frac{1}{6}&\frac{1}{6}&0\end{bmatrix}$$

$$M^2(G_m,N)=\begin{bmatrix}0&0&0&0&0&\frac{1}{6}\\0&0&0&0&0&0\\0&0&0&0&0&\frac{1}{10}\\0&0&0&0&0&\frac{1}{12}\\0&0&0&0&0&\frac{1}{14}\\\frac{1}{8}&0&\frac{1}{16}&\frac{1}{16}&\frac{1}{16}&\frac{3}{8}\end{bmatrix}$$

Table 6. Values of amino acid-based (local)bilinear indices for hetero-dimer SKEERN.

Amino-acid (AA)	*Local Non-Stochastic Bilinear Indices*		
	$b_{0L}(\bar{x}_m, \bar{y}_m)$	$b_{1L}(\bar{x}_m, \bar{y}_m)$	$b_{2L}(\bar{x}_m, \bar{y}_m)$
Ser (S)	-3.1948	-0.8104	-13.0522
Lys (K)	4.0044	6.1215	28.6812
Glu (E)	1.2012	3.9264	5.8605
Glu (E)	1.2012	7.3033	10.3029
Arg (R)	7.2576	10.71	43.578
Asn (N)	4.669	13.3352	23.4674
Hetero-Dimer (SKEERN)	**15.1386**	**40.586**	**98.8378**
Amino Acid (AA)	*Local Stochastic Bilinear Indices*		
	$^{s}b_{0L}(\bar{x}_m, \bar{y}_m)$	$^{s}b_{1L}(\bar{x}_m, \bar{y}_m)$	$^{s}b_{2L}(\bar{x}_m, \bar{y}_m)$
Ser (S)	-3.1948	0.37176667	-2.04034833
Lys (K)	4.0044	2.6327	4.27309429
Glu (E)	1.2012	1.8709	1.08062179
Glu (E)	1.2012	3.4534	1.66443036
Arg (R)	7.2576	4.6284	6.24537857
Asn (N)	4.669	4.81723333	3.34964405
Hetero-Dimer (SKEERN)	**15.1386**	**17.7744**	**14.5728207**

Table 7. Experimental and calculated values of melting temperature (t_m) obtained by equation **23**.

Mutant	Obs.[a]	Cal.[b]	Res.[c]	Res_{CV}[d]	Mutant	Obs.[a]	Cal.[b]	Res.[c]	Res_{CV}[d]
1PA8-st6	74.1		*outlier*		25EA43-st6	56.1	51.7	4.4	4.9
2SA35-st6	63.4	59.1	4.3	5.0	26EA28-st11	55.7	56.3	-0.6	-0.7
3NA34-st11	63.0	55.6	7.4	8.5	27MA7-st6	55.5	53.8	1.7	1.9
4NA11-st6	62.1	59.4	2.7	3.4	28DA20-st6	55.3	60.0	-4.7	-6.0
5QA39-st11	61.4	56.3	5.1	5.5	29IA51-st11	50.9	50.7	0.2	0.3
6GA52-st11	60.9	63.3	-2.4	-3.2	30GA49-st11	48.7	51.3	-2.6	-3.7
7KA6-st6	59.6	62.4	-2.8	-3.2	31LA19-st6	48.3	46.2	2.1	2.4
8RA16-st6	59.5	57.2	2.3	2.7	32GA30-st11	47.9	45.7	2.2	2.6
9VA25-st6	59.3	56.1	3.2	3.4	33RA50-st11	47.9	46.6	1.3	1.5
10MA4-st6	59.2	60.1	-0.9	-1.1	34KA47-st11	47.2	47.1	0.1	0.1
11Arc-st6	59.0	60.2	-1.2	-1.3	35PA15-st11	46.6	47.9	-1.3	-1.5
12EA27-st6	58.8	58.7	0.1	0.1	36SA44-st11	46.3	43.2	3.1	3.9
13KA2-st6	58.7	58.3	0.4	0.5	37NA29-st11	45.3	45.7	-0.4	-0.4
14QA9-st6	58.4	60.2	-1.8	-2.0	38VA33-st11	44.1	48.8	-4.7	-5.0
15GA3-st6	58.1	61.0	-2.9	-3.1	39EA48-st11	43.2	46.3	-3.1	-3.4
16MA1-st6	58.0	54.8	3.2	3.5	40LA12-st11	42.3	40.6	1.7	1.9
17Arc-st11	57.9	53.5	4.4	4.8	41FA10-st6	40.6	47.3	-6.7	-8.3
18SA5-st6	57.5	60.7	-3.2	-3.4	42LA21-st11	39.6	39.2	0.4	0.5
19RA13-st6	57.3	56.9	0.4	0.5	43RA31-st11	37.1	41.6	-4.5	-5.0
20KA46-st11	57.1	54.5	2.6	2.9	44MA42-st11	35.6	42.3	-6.7	-7.5
21EA17-st6	57.0	62.8	-5.8	-6.5	45SA32-st11	33.5		*outlier*	
22VA18-st6	56.9	52.2	4.7	5.2	46YA38-st11	33.0	39.1	-6.1	-7.2
23RA23-st11	56.7	49.5	7.2	7.6	47WA14-st11	31.5	27.3	4.2	7.8
24KA24-st11	56.3	59.8	-3.5	-3.9	48RA40-st11	31.2	34.7	-3.5	-6.8

[a]Experimental melting temperature t_m ^{0}C.[34] Proteins are arranged in order of decreasing t_m, Mutants 49–53 (VA22-st11, EA36-st11, IA37-st11,VA41-st11 and FA45-st11) were extracted in the QSAR study due to its nonaccurate t_m values (<20 ^{0}C), which is not useful for MLR analysis. The st6 and st11 refer to C-terminal sequences of the mutant proteins.[34]

[b]Calculated t_m values by the Eq. **23**.

[c]Residual: t_m (Obs.) - t_m (Cal.).

[d]Residual by LOO cross-validation procedures (deleted residual).

Table 8. Experimental and calculated values of melting temperature (t_m) obtained by equation **24**.

Mutant	Obs.[a]	Cal.[b]	Res.[c]	Res_{CV}[d]	Mutant	Obs.[a]	Cal.[b]	Res.[c]	Res_{CV}[d]
1PA8-st6	74.1		*outlier*		25EA43-st6	56.1	56.1	0.0	0.0
2SA35-st6	63.4	59.0	4.4	4.8	26EA28-st11	55.7	53.0	2.7	3.8
3NA34-st11	63.0	55.9	7.1	9.0	27MA7-st6	55.5	56.3	-0.8	-0.9
4NA11-st6	62.1	57.0	5.1	7.0	28DA20-st6	55.3	62.9	-7.6	-10.3
5QA39-st11	61.4	52.9	8.5	9.1	29IA51-st11	50.9	49.7	1.2	1.3
6GA52-st11	60.9	60.7	0.2	0.3	30GA49-st11	48.7	55.5	-6.8	-9.9
7KA6-st6	59.6	59.8	-0.2	-0.2	31LA19-st6	48.3	47.6	0.7	0.8
8RA16-st6	59.5	61.1	-1.6	-1.9	32GA30-st11	47.9	43.9	4.0	4.7
9VA25-st6	59.3	56.9	2.4	2.7	33RA50-st11	47.9	53.6	-5.7	-7.7
10MA4-st6	59.2	52.4	6.8	7.5	34KA47-st11	47.2	51.9	-4.7	-5.0
11Arc-st6	59.0	59.6	-0.6	-0.6	35PA15-st11	46.6	50.0	-3.4	-4.1
12EA27-st6	58.8	60.8	-2.0	-2.3	36SA44-st11	46.3	47.0	-0.7	-0.7
13KA2-st6	58.7	56.3	2.4	3.0	37NA29-st11	45.3	42.6	2.7	3.0
14QA9-st6	58.4	61.9	-3.5	-3.9	38VA33-st11	44.1	47.9	-3.8	-4.1
15GA3-st6	58.1	60.0	-1.9	-2.0	39EA48-st11	43.2	47.8	-4.6	-5.3
16MA1-st6	58.0	59.1	-1.1	-1.2	40LA12-st11	42.3	37.4	4.9	6.5
17Arc-st11	57.9	52.7	5.2	6.0	41FA10-st6	40.6	43.6	-3.0	-6.2
18SA5-st6	57.5	56.8	0.7	0.7	42LA21-st11	39.6	39.8	-0.2	-0.3
19RA13-st6	57.3	60.9	-3.6	-4.4	43RA31-st11	37.1	37.4	-0.3	-0.5
20KA46-st11	57.1	53.8	3.3	3.5	44MA42-st11	35.6	40.8	-5.2	-5.7
21EA17-st6	57.0	58.4	-1.4	-1.6	45SA32-st11	33.5		*outlier*	
22VA18-st6	56.9	53.9	3.0	3.2	46YA38-st11	33.0	33.1	-0.1	-0.1
23RA23-st11	56.7	51.2	5.5	6.4	47WA14-st11	31.5	38.0	-6.5	-8.2
24KA24-st11	56.3	56.7	-0.4	-0.5	48RA40-st11	31.2	32.6	-1.4	-1.9

[a]Experimental melting temperature. t_m.^{0}C.[34] Proteins are arranged in order of decreasing t_m, Mutants 49–53 (VA22-st11, EA36-st11, IA37- st11, VA41- st11, and FA45-st11) were extracted in the QSAR study due to its nonaccurate t_m values (<20 ^{0}C), which is not useful for MLR analysis. The st6 and st11 refer to C-terminal sequences of the mutant proteins.[34]

[b]Calculated t_m values by the Eq. **24.**

[c]Residual: t_m (Obs.) - t_m (Cal.).

[d]Residual by LOO cross-validation procedures (deleted residual).

Table 9. Results of the non-stochastic bilinear indices-driven ADL models of the Arc A-mutants in the training and test set.

Mutant	ΔP%[b]	P(H)[c]	P(P)[c]	Mutant	ΔP%[b]	P(H)[c]	P(P)[c]
Mutants with near wild type stability (H)				*Mutants with reduced stability(P)*			
1PA8-st6[a]	99.95	1.00	0.00	29IA51-st11	-99.11	0.00	1.00
2SA35-st6	92.63	0.96	0.04	30GA49-st11[a]	-59.42	0.20	0.80
3NA34-st11	94.96	0.97	0.03	31LA19-st6	-4.14	0.48	0.52
4NA11-st6[a]	99.96	1.00	0.00	32GA30-st11	-98.66	0.01	0.99
5QA39-st11	99.60	1.00	0.00	33RA50-st11	-77.55	0.11	0.89
6GA52-st11	9.67	0.55	0.45	34KA47-st11	-34.15	0.33	0.67
7KA6-st6[a]	100.00	1.00	0.00	35PA15-st11[a]	-63.06	0.18	0.82
8RA16-st6	99.97	1.00	0.00	36SA44-st11	-99.98	0.00	1.00
9VA25-st6	98.45	0.99	0.01	37NA29-st11	-99.90	0.00	1.00
10MA4-st6	99.50	1.00	0.00	38VA33-st11	-99.82	0.00	1.00
11Arc-st6[a]	99.99	1.00	0.00	39EA48-st11	-16.56	0.42	0.58
12EA27-st6	99.67	1.00	0.00	40LA12-st11	-99.82	0.00	1.00
13KA2-st6	100.00	1.00	0.00	*41FA10-st6[a]	76.85	0.88	0.12
14QA9-st6	99.98	1.00	0.00	42LA21-st11	-99.97	0.00	1.00
15GA3-st6	99.98	1.00	0.00	43RA31-st11	-99.80	0.00	1.00
16MA1-st6[a]	99.83	1.00	0.00	44MA42-st11	-97.57	0.01	0.99
17Arc-st11	62.49	0.81	0.19	45SA32-st11[a]	-37.11	0.31	0.69
18SA5-st6	99.99	1.00	0.00	46YA38-st11	-85.72	0.07	0.93
19RA13-st6	100.00	1.00	0.00	47WA14-st11	-98.49	0.01	0.99
20KA46-st11	99.23	1.00	0.00	48RA40-st11	-100.00	0.00	1.00
21EA17-st6[a]	100.00	1.00	0.00	49VA22-st11	-97.68	0.01	0.99
22VA18-st6	91.02	0.96	0.04	50EA36-st11[a]	-99.64	0.00	1.00
23RA23-st11	12.81	0.56	0.44	51IA37-st11	-99.99	0.00	1.00
24KA24-st11	97.78	0.99	0.01	52VA41-st11	-99.96	0.00	1.00
25EA43-st6	99.72	1.00	0.00	53FA45-st11	-100.00	0.00	1.00
26EA28-st11[a]	43.96	0.72	0.28				
27MA7-st6	99.26	1.00	0.00				
28DA20-st6	99.90	1.00	0.00				

*Mutants that are misclassified by model **21**.

[a]Compounds in the test set.

[b]ΔP% = [P(H-group) - P(P-group)] ×100

[c]Percentage of probability with which the mutants is predicted as reduced stability/near wild-type stability mutants, respectively.

Table 10. Results of the stochastic bilinear indices-driven ADL models of the Arc A-mutants in the training and test set.

Mutant	ΔP%[a]	P(H)[b]	P(P)[c]	Mutant	ΔP%[a]	P(H)[b]	P(P)[c]
Mutants with near wild type stability				*Mutants with reduced stability*			
1PA8-st6[a]	90.81	0.95	0.05	29IA51-st11	-99.82	0.00	1.00
2SA35-st6	99.33	1.00	0.00	30GA49-st11[a]	-97.78	0.01	0.99
3NA34-st11	85.37	0.93	0.07	31LA19-st6	-23.61	0.38	0.62
4NA11-st6[a]	82.75	0.91	0.09	32GA30-st11	-99.40	0.00	1.00
5QA39-st11	83.47	0.92	0.08	33RA50-st11	-99.13	0.00	1.00
6GA52-st11	5.76	0.53	0.47	*34KA47-st11	47.28	0.74	0.26
7KA6-st6[a]	99.67	1.00	0.00	35PA15-st11[a]	-37.09	0.31	0.69
8RA16-st6	100.00	1.00	0.00	36SA44-st11	-85.82	0.07	0.93
9VA25-st6	66.11	0.83	0.17	37NA29-st11	-95.25	0.02	0.98
10MA4-st6	13.62	0.57	0.43	38VA33-st11	-98.80	0.01	0.99
11Arc-st6[a]	100.00	1.00	0.00	39EA48-st11	-94.11	0.03	0.97
12EA27-st6	98.78	0.99	0.01	40LA12-st11	-99.99	0.00	1.00
13KA2-st6	99.10	1.00	0.00	41FA10-st6[a]	-89.82	0.05	0.95
14QA9-st6	99.38	1.00	0.00	42LA21-st11	-99.85	0.00	1.00
15GA3-st6	96.73	0.98	0.02	43RA31-st11	-99.41	0.00	1.00
16MA1-st6[a]	87.80	0.94	0.06	44MA42-st11	-98.86	0.01	0.99
17Arc-st11	99.69	1.00	0.00	45SA32-st11[a]	-81.42	0.09	0.91
18SA5-st6	99.71	1.00	0.00	46YA38-st11	-96.44	0.02	0.98
19RA13-st6	99.99	1.00	0.00	47WA14-st11	-96.27	0.02	0.98
20KA46-st11	37.83	0.69	0.31	48RA40-st11	-27.72	0.36	0.64
21EA17-st6[a]	99.79	1.00	0.00	49VA22-st11	-98.63	0.01	0.99
22VA18-st6	73.50	0.87	0.13	*50EA36-st11[a]	57.60	0.79	0.21
23RA23-st11	95.59	0.98	0.02	51IA37-st11	-98.60	0.01	0.99
24KA24-st11	79.13	0.90	0.10	52VA41-st11	-97.23	0.01	0.99
25EA43-st6	99.73	1.00	0.00	53FA45-st11	-99.81	0.00	1.00
26EA28-st11[a]	94.00	0.97	0.03				
27MA7-st6	85.08	0.93	0.07				
28DA20-st6	100.00	1.00	0.00				

*Mutants that are misclassified by model **22**.

[a]Compounds in the test set.

[b]ΔP% = [P(H-group) - P(P-group)] × 100

[c]Percentage of probability with which the mutants is predicted as reduced stability/near wild-type stability mutants, respectively.

Table 11.Experimental and calculated values of melting temperature (t_m) obtained by Eqs. **25** and **26**.

Mutant	Obs.[a]	Cal.[b]	Res.[c]	Mutant	Obs.[a]	Cal.[b]	Res.[c]
1PA8-st6	74.1	*outlier*		25EA43-st6	56.1	56.4	-0.3
2SA35-st6	63.4	58.8	4.6	26EA28-st11	55.7	58.3	-2.6
3NA34-st11	63.0	58.4	4.6	27MA7-st6	55.5	57.8	-2.3
4NA11-st6	62.1	59.0	3.1	28DA20-st6	55.3	57.7	-2.4
5QA39-st11	61.4	58.9	2.5	29IA51-st11	50.9	49.0	1.9
6GA52-st11	60.9	60.3	0.6	30GA49-st11	48.7	50.1	-1.4
7KA6-st6	59.6	59.2	0.4	31LA19-st6	48.3	43.6	4.7
8RA16-st6	59.5	57.0	2.5	32GA30-st11	47.9	45.0	2.9
9VA25-st6	59.3	58.3	1.0	33RA50-st11	47.9	44.6	3.3
10MA4-st6	59.2	58.4	0.8	34KA47-st11	47.2	44.6	2.6
11Arc-st6	59.0	59.0	0.0	35PA15-st11	46.6	46.5	0.1
12EA27-st6	58.8	58.3	0.5	36SA44-st11	46.3	41.6	4.7
13KA2-st6	58.7	58.7	0.0	37NA29-st11	45.3	44.5	0.8
14QA9-st6	58.4	59.1	-0.7	38VA33-st11	44.1	47.2	-3.1
15GA3-st6	58.1	59.1	-1.0	39EA48-st11	43.2	44.9	-1.7
16MA1-st6	58.0	58.0	0.0	40LA12-st11	42.3	40.7	1.6
17Arc-st11	57.9	58.8	-0.9	41FA10-st6	40.6	44.4	-3.8
18SA5-st6	57.5	59.0	-1.5	42LA21-st11	39.6	39.5	0.1
19RA13-st6	57.3	58.0	-0.7	43RA31-st11	37.1	40.1	-3.0
20KA46-st11	57.1	58.3	-1.2	44MA42-st11	35.6	41.3	-5.7
21EA17-st6	57.0	59.1	-2.1	45SA32-st11	33.5	*outlier*	
22VA18-st6	56.9	57.8	-0.9	46YA38-st11	33.0	38.2	-5.2
23RA23-st11	56.7	57.0	-0.3	47WA14-st11	31.5	27.8	3.7
24KA24-st11	56.3	59.8	-3.5	48RA40-st11	31.2	33.7	-2.5

[a]Experimental melting temperature. t_m.^{0}C.[34] Proteins are arranged in order of decreasing t_m. Mutants 49–53 (VA22-st11, EA36-st11, IA37- st11,VA41- st11and FA45-st11) were extracted in the QSAR study due to its nonaccurate t_m values (<20 ^{0}C), which is not useful for Piecewise method. The st6 and st11 refer to C-terminal sequences of the mutant proteins.[34]

[b]Calculated t_m values by the nonlinear model Eqs. **25** and **26**.

[c]Residual: t_m (Obs.) - t_m (Cal.).

Table 12.Experimental and calculated values of melting temperature (t_m) obtained by eqs. **27** and **28**.

Mutant	Obs.[a]	Cal.[b]	Res.[c]	Mutant	Obs.[a]	Cal.[b]	Res.[c]
1PA8-st6	74.1	*outlier*		25EA43-st6	56.1	58.3	-2.2
2SA35-st6	63.4	58.3	5.1	26EA28-st11	55.7	56.7	-1.0
3NA34-st11	63.0	59.4	3.6	27MA7-st6	55.5	58.2	-2.7
4NA11-st6	62.1	60.0	2.1	28DA20-st6	55.3	57.9	-2.6
5QA39-st11	61.4	58.2	3.2	29IA51-st11	50.9	46.3	4.6
6GA52-st11	60.9	59.6	1.3	30GA49-st11	48.7	51.6	-2.9
7KA6-st6	59.6	58.3	1.3	31LA19-st6	48.3	43.5	4.8
8RA16-st6	59.5	56.9	2.6	32GA30-st11	47.9	43.5	4.4
9VA25-st6	59.3	59.1	0.2	33RA50-st11	47.9	51.5	-3.6
10MA4-st6	59.2	58.1	1.1	34KA47-st11	47.2	48.7	-1.5
11Arc-st6	59.0	58.9	0.1	35PA15-st11	46.6	46.6	0.0
12EA27-st6	58.8	58.3	0.5	36SA44-st11	46.3	45.2	1.1
13KA2-st6	58.7	57.9	0.8	37NA29-st11	45.3	42.0	3.3
14QA9-st6	58.4	59.5	-1.1	38VA33-st11	44.1	45.6	-1.5
15GA3-st6	58.1	58.8	-0.7	39EA48-st11	43.2	43.5	-0.3
16MA1-st6	58.0	58.7	-0.7	40LA12-st11	42.3	39.2	3.1
17Arc-st11	57.9	58.7	-0.8	41FA10-st6	40.6	43.0	-2.4
18SA5-st6	57.5	58.6	-1.1	42LA21-st11	39.6	38.3	1.3
19RA13-st6	57.3	58.2	-0.9	43RA31-st11	37.1	36.1	1.0
20KA46-st11	57.1	58.2	-1.1	44MA42-st11	35.6	38.7	-3.1
21EA17-st6	57.0	59.1	-2.1	45SA32-st11	33.5	*outlier*	
22VA18-st6	56.9	58.1	-1.2	46YA38-st11	33.0	33.3	-0.3
23RA23-st11	56.7	57.2	-0.5	47WA14-st11	31.5	36.5	-5.0
24KA24-st11	56.3	59.6	-3.3	48RA40-st11	31.2	34.1	-2.9

[a]Experimental melting temperature. t_m.^{0}C.[34] Proteins are arranged in order of decreasing t_m. Mutants 49–53 (VA22-st11, EA36-st11, IA37-st11,VA41-st11, and FA45-st11) were extracted in the QSAR study due to its nonaccurate t_m values (<20 ^{0}C), which is not useful for Piecewise method. The st6 and st11 refer to C-terminal sequences of the mutant proteins.[34]

[b]Calculated t_m values by the nonlinear model Eqs. **27** and **28**.

[c]Residual: t_m (Obs.) - t_m (Cal.).

Table 13. Comparison between LDA statistical parameters from protein's bilinear indices classification models with other reported *´in silico*' methods.

Methods[a]	Accuracy (%)	%Nwt[b]	%RS[b]	%NC[b]	N	λ Wilks	F	p-level	MCC	Model Descriptors	Eq/Ref
Non-Stochastic Protein Bilinear indices	100	100	100	0.0	41	0.24	28.08	<0.0001	1.00	$Class$ =-45,329-5.00x10^{-3} $^{Z1\text{-}ISA}b_0(\bar{x}_m,\bar{y}_m)$ - 1.00x10^{-3} $^{Z2\text{-}Z3}b_6(\bar{x}_m,\bar{y}_m)$ +2.00 x10^{-3} $^{Z2\text{-}HPI}b_5(\bar{x}_m,\bar{y}_m)$-0,435$^{ECI\text{-}PI}b_2(\bar{x}_m,\bar{y}_m)$	
Stochastic Protein Bilinear indices	97.56	100	95.00	0.0	41	0.29	21.61	<0.0001	0.95	$Class$ = 24.797-5.00x10^{-3} $^{Z1\text{-}ISA}$ $^{s}b_2(\bar{x}_m,\bar{y}_m)$ -53,074$^{ECI\text{-}HPIs}b_0(\bar{x}_m,\bar{y}_m)$-0,465$^{Z2\text{-}ECI}$ $^{s}b_1(\bar{x}_m,\bar{y}_m)$ - 0,152 $^{Z2\text{-}HPI}$ $^{s}b_6(\bar{x}_m,\bar{y}_m)$	
Linear Indices	97.56	95.23	100	0.0	41	0.31	15.25	<0.0001	0.95		[47]
Quadratic Indices	85.4	85.0	85.7	0.0	41	0.47	9.89	<0.0001	0.71		[44]
Protein Stochastic Moments	81.13	71.4	92.0	-	53	0.63	14.5	<0.001	-		[87]
ξ_1	81.1	71.4	92.0	-	53	0.63	29.57	<0.001	-		[29]
$\Delta\theta_0$	81.1	71.4	92.0	0.0	53	0.56	39.05	0.00	0.64		[30]
D-Fire	76.9	92.9	58.3	3.8	53	0.79	13.9	0.00	0.55		[30]
Surface	70.7	63.6	78.9	22.6	53	0.85	8.8	0.00	0.43		[30]
Volume	62.3	53.6	72.0	0.0	53	0.92	4.2	0.00	0.26		[30]
Log P	59.0	80.8	15.4	26.4	53	0.99	0.5	0.5	0.05		[30]
Refractivity	60.0	77.3	38.9	24.5	53	0.97	1.8	0.2	0.18		[30]

Validation Method

Methods[a]	Validation method[c]	Accuracy (test set)[d]	%TL-25%-O [b]	D^2	F	P (F) - level	MCC
Non-Stochastic Protein Bilinear indices	i	91.67	-	11.88	8.08	<0.0001	0.84
Stochastic Protein Bilinear indices	i	91.67	-	9.14	1.61	<0.0001	0.84
Linear Indices	i	91.67	-	8.72	5.25	<0.0001	0.84
Quadratic Indices	i	91.67	-	4.40	9.89	<0.0001	0.84
Protein Stochastic Moments	-	-	-				
ξ_1	-	-	-				
$\Delta\theta_0$	ii	N	79.5				
D-Fire	ii	N	71.8				
Surface	ii	N	61.5				
Volume	ii	N	56.4				
Log P	ii	N	48.7				
Refractivity	ii	N	61.5				

[a] Non-stochastic and Stochastic Bilinear Indices are reported in this work; $\Delta\theta_0$, D-Fire, surface, volume, log P, and refractivity are reported by R de Armas et al.;[30] Protein stochastic moments are published in [87] and ξ_1 in [29].

[b] Parameters verifying model quality: %Nwt, %RS, %NC, %TL-25%-O are the near wild-type group, reduced-stability group, nonclassified, and total after leave-25%-out percentages of good classification.

[c] Validation methods are: (i) test set and (ii) leave-25%-out.

[d] Test set of 12 A-mutants of the Arc repressor.

Table 14. Comparison between LMR parameters of Protein's Bilinear Indices and other ***TOMOCOMD-CARDD*** reported methods.

Methods	Linear Multiple Regression parameters						
	R	R^2	s	q^2	s_{cv}	F	p-level
Nonstochastic Protein's Bilinear Indices	0.91	0.83	3.9	0.73	4.50	33.08	<0.0001
Stochastic Protein's Bilinear Indices	0.90	0.81	4.19	0.64	5.01	28.24	<0.0001
Protein's Linear Indices[47]	0.90	0.81	4.29	0.72	4.79	26.48	<0.0001
Protein's Quadratic Indices[44]	0.85	0.72	5.64	0.55	6.24	9.04	<0.0001

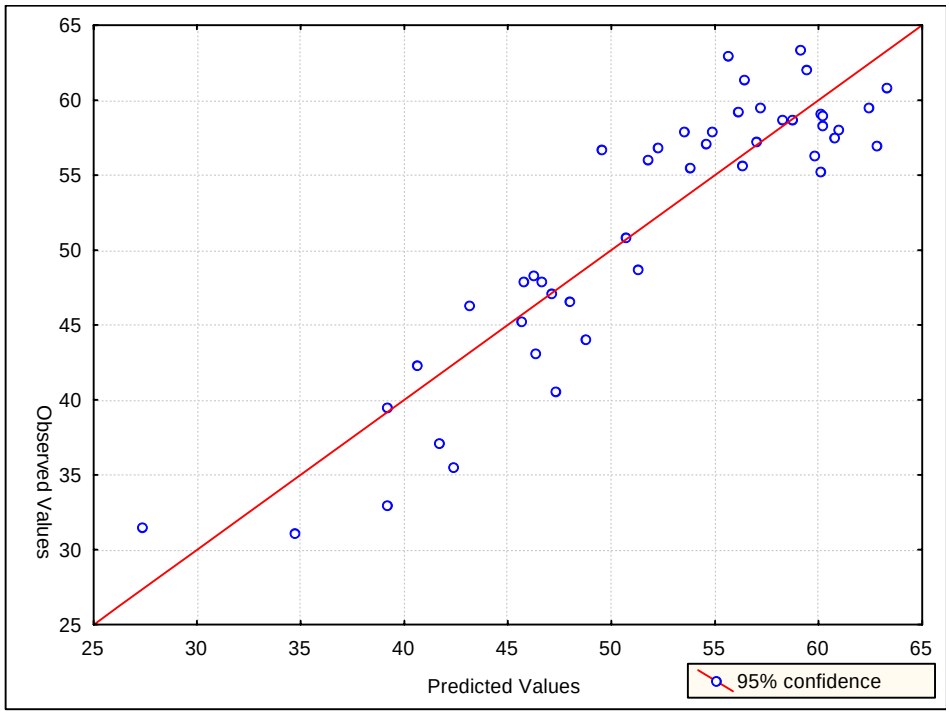


Figure 1. Linear correlations of observed versus calculated melting point according to the model obtained from non-stochastic bilinear indices (Eq. **23**).

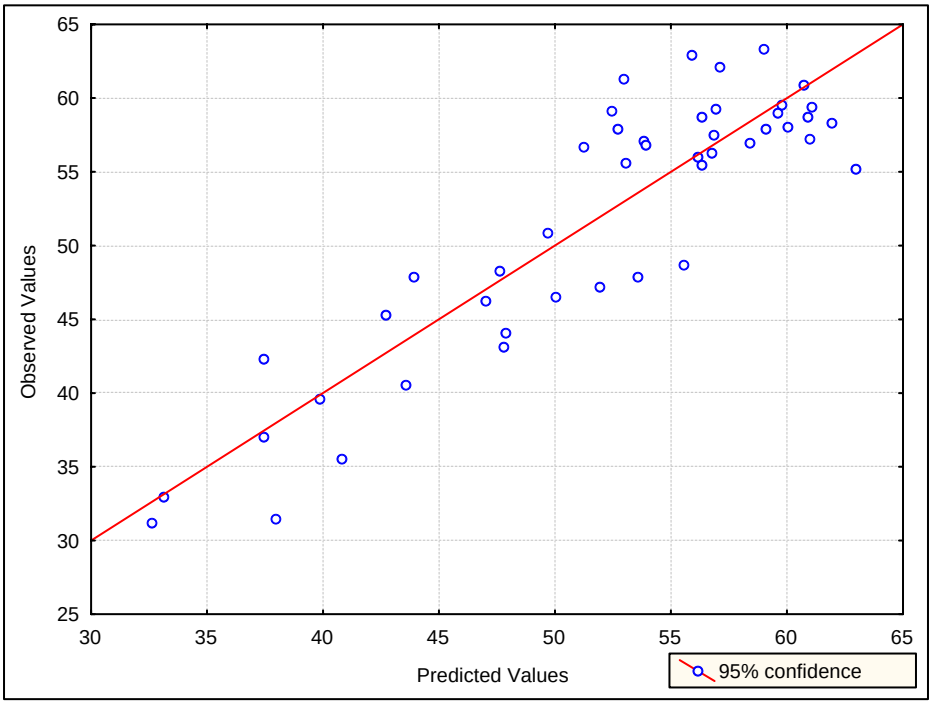
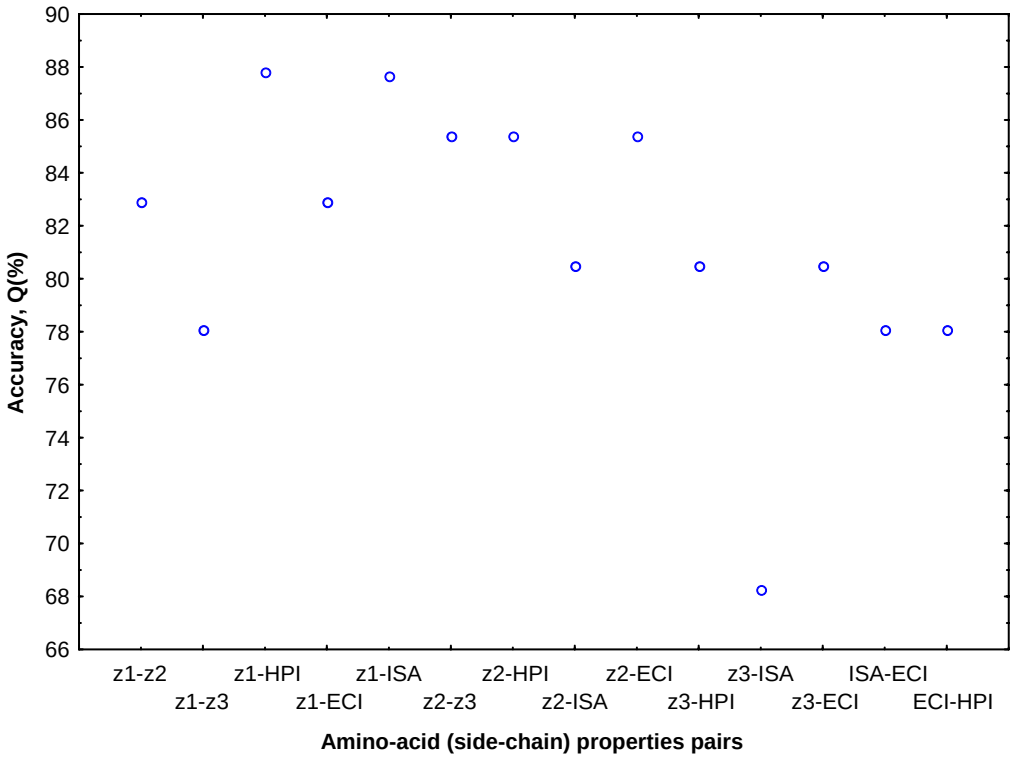
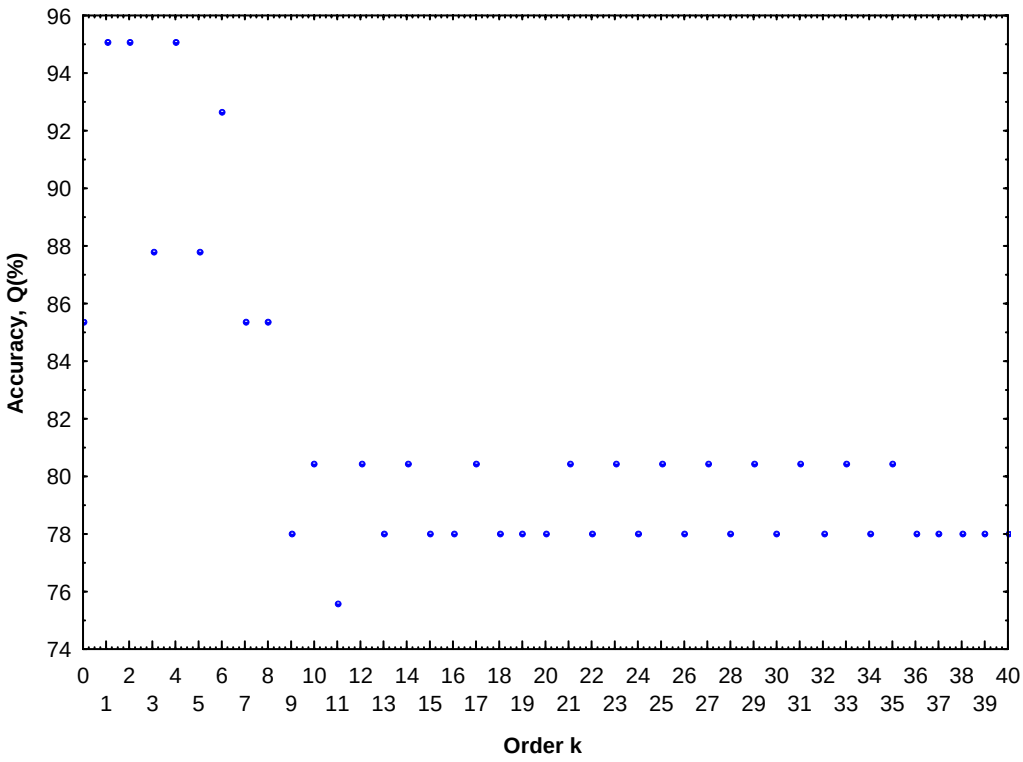


Figure 2. Linear correlations of observed versus calculated melting point according to the model obtained from stochastic bilinear indices (Eq. **24**).



101 **Figure 3**. Dependence of global good classification (accuracy) between t_m (two-class)
102 and the protein bilinear indices calculated by using different amino-acid weights, which
103 was composed by the pairs-combination of six amino-acid side-chain properties.



104
105 **Figure 4.** Dependence of global good classification (accuracy) between t_m (two-class)
106 and the protein bilinear indices calculated at different orders k (k = 0-40).