Available online at www.sciencedirect.com
SCIENCE DIRECT®
Bioorganic & Medicinal Chemistry 13 (2005) 3003–3015
Bioorganic & Medicinal Chemistry

Protein linear indices of the 'macromolecular pseudograph α-carbon atom adjacency matrix' in bioinformatics. Part 1: Prediction of protein stability effects of a complete set of alanine substitutions in Arc repressor

Yovani Marrero-Ponce,[a,b,*] Ricardo Medina-Marrero,[c] Juan A. Castillo-Garit,[b,d] Vicente Romero-Zaldivar,[e] Francisco Torrens[f] and Eduardo A. Castro[g]

[a]*Department of Pharmacy, Faculty of Chemical-Pharmacy, Central University of Las Villas, Santa Clara, 54830 Villa Clara, Cuba*
[b]*Department of Drug Design, Chemical Bioactive Center, Central University of Las Villas, Santa Clara, 54830 Villa Clara, Cuba*
[c]*Department of Microbiology, Chemical Bioactive Center, Central University of Las Villas, Santa Clara, 54830 Villa Clara, Cuba*
[d]*Applied Chemistry Research Center, Central University of Las Villas, Santa Clara 54830, Villa Clara, Cuba*
[e]*Faculty of Informatics, University of Cienfuegos, Cienfuegos 55500, Cuba*
[f]*Institut Universitari de Ciència Molecular, Universitat de València, Dr. Moliner 50, E-46100 Burjassot (València), Spain*
[g]*INIFTA, División Química Teórica, Suc.4, C.C. 16, La Plata 1900 Buenos Aires, Argentina*

Received 22 December 2004; revised 28 January 2005; accepted 31 January 2005

Abstract—A novel approach to bio-macromolecular design from a linear algebra point of view is introduced. A protein's total (whole protein) and local (one or more amino acid) linear indices are a new set of bio-macromolecular descriptors of relevance to protein QSAR/QSPR studies. These amino-acid level biochemical descriptors are based on the calculation of linear maps on $\Re^n[f_k(x_{mi}) : \Re^n \rightarrow \Re^n]$ in canonical basis. These bio-macromolecular indices are calculated from the k^{th} power of the macromolecular pseudograph α-carbon atom adjacency matrix. Total linear indices are linear functional on $\Re^n$. That is, the k^{th} total linear indices are linear maps from $\Re^n$ to the scalar $\Re[f_k(x_m) : \Re^n \rightarrow \Re]$. Thus, the k^{th} total linear indices are calculated by summing the amino-acid linear indices of all amino acids in the protein molecule. A study of the protein stability effects for a complete set of alanine substitutions in the Arc repressor illustrates this approach. A quantitative model that discriminates near wild-type stability alanine mutants from the reduced-stability ones in a training series was obtained. This model permitted the correct classification of 97.56% (40/41) and 91.67% (11/12) of proteins in the training and test set, respectively. It shows a high Matthews correlation coefficient (MCC = 0.952) for the training set and an MCC = 0.837 for the external prediction set. Additionally, canonical regression analysis corroborated the statistical quality of the classification model (R_{canc} = 0.824). This analysis was also used to compute biological stability canonical scores for each Arc alanine mutant. On the other hand, the linear piecewise regression model compared favorably with respect to the linear regression one on predicting the melting temperature (t_m) of the Arc alanine mutants. The linear model explains almost 81% of the variance of the experimental t_m (R = 0.90 and s = 4.29) and the LOO press statistics evidenced its predictive ability (q^2 = 0.72 and s_{cv} = 4.79). Moreover, the *TOMOCOMD-CAMPS* method produced a linear piecewise regression (R = 0.97) between protein backbone descriptors and t_m values for alanine mutants of the Arc repressor. A break-point value of 51.87 °C characterized two mutant clusters and coincided perfectly with the experimental scale. For this reason, we can use the linear discriminant analysis and piecewise models in combination to classify and predict the stability of the mutant Arc homodimers. These models also permitted the interpretation of the driving forces of such folding process, indicating that topologic/topographic protein 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; Protein linear indices; QSAR.
* Corresponding author. Tel.: +53 42 281192/281473; fax: +53 42 281130/281455; e-mail addresses: yovanimp@qf.uclv.edu.cu; ymarrero77@yahoo.es

What are the Limits of Applicability for Graph Theoretic Descriptors in QSPR/QSAR?
[Estrada, E and González, H. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 75–84.]

0968-0896/$ - see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.bmc.2005.01.062

1. Introduction

Anfinsen's experiment with ribonuclease A and *staphylococcal* nuclease discovered that amino acid sequence of these small proteins encode their final folded structure and also encode the information on how to get to the structures.[1,2] However, the 'folding problem (prediction of the three-dimensional structure of a protein from its amino-acid sequence)' still remains as one of the greatest unsolved problems of protein science. The folding problem is very important due to the large number of genome sequences completed in recent years. This fact has provoked a large gap between the sharply increasing number of protein sequences entering into data banks and the slow accumulation of known structures. Thus, predicting the spatial structure based on a given protein primary-sequence information could play a significant role in conjunction with experimental methods.[3]

The major constituent of proteins is an unbranched polypeptide chain consisting of L-α-amino acids linked by amide bonds between the α-carboxyl group of one residue and the α-amino group of the next. The sequence of the amino acids defines the primary structure.[4–9] As previously outlined, the genetically encoded sequence of a protein determines its three-dimensional structure.[1,2,4–9] That is to say, if the side chain of each amino acid within a protein is removed, the secondary structure of the protein is obtained. It is constructed around planar units of the peptide bonds. A closer examination reveals regions where the secondary structure is organized into repetitive and regular elements.

Afterwards, the side chains can be added back to the backbone, and it is then seen how the ternary structure of the proteins is formed by packing the regular elements of the secondary structure through their side chains. For this reason, the structure of each protein can be expressed in a quantitative way by side-chain amino-acid properties. Subsequently, Charton and Charton determined the dependence of protein conformation upon the side-chain structure of the amino-acid residues using Chou-Fasman parameters.[10]

In an other approach about structure–activity studies, Hellberg et al. developed the so-called principal properties or *z*-values.[11] This peptide QSAR methodology is based on a parametrization of each amino acid occurring in a peptide chain with three *z*-values, which are linear combinations of the original measured variables. These values are proposed to be related to hydrophilicity, bulk, and electronic properties. The principal properties have been successfully used to seek peptide QSARs.[11–13] Other descriptors used in peptide QSAR studies have been derived from the side-chain surface area and the atomic charges of amino acids.[14]

Hydrophobicity (or hydrophilicity) plots have the goal of predicting membrane-spanning segments (highly hydrophobic) or regions that are likely exposed on the surface of proteins (hydrophilic domains) and therefore potentially antigenic. In this context, several hydrophobic scales have been developed, most of which were derived from experimental studies on partitioning of peptides in apolar and polar solvents.[15,16]

On the other hand, most of the properties of very large systems, as bio-macromolecules and supramolecular complexes, can be assessed with simplified models. For example, in proteins, amino acid residues can be depicted using a lower level representation, that is, two or three pseudo-atoms rather than by an all-atom representation.[17,18] The advantage of using nonatomic representation is, however, not limited to the increase of the speed of computations. Simplified representations of protein geometry have also been used by many groups to reduce sensitivity to small perturbations in conformation, for example, when docking a ligand versus a receptor.[19,20] Cherfils et al.[19] replaced amino acid residues with spheres of varying sizes and performed docking to maximize the buried surface area.

In this sense, our research group has recently introduced the novel computer-aided molecular design scheme *TOMOCOMD-CARDD* (acronym of *TO*pological *MO*lecular *COM*puter *D*esign-*C*omputer *A*ided '*R*ational' *D*rug *D*esign).[21–23] This method has been developed to generate molecular descriptors based on the linear algebra theory. The approach describes changes in the electron distribution with time throughout the molecular backbone. It has been successfully employed in QSPR/QSAR studies,[24–28] including studies related to nucleic acid–drug interactions.[29] One of the applications involved the prediction of the anthelmintic activity of novel drugs.[24,30] More recently, the *TOMOCOMD-CARDD* approach has been applied to the fast-track experimental discovery of novel paraphistomicide drug-like compounds.[31] Codification of chirality and other 3D structural features constitutes another advantage of this method.[32] The latter opportunity has allowed the description of the significance interpretation and the comparison to other molecular descriptors.[23,33]

The main aim of this paper is to propose an extended *TOMOCOMD* approach to account for protein structure. In the present study, we propose a total and local definition of protein linear indices of the 'macromolecular pseudograph α-carbon atom adjacency matrix'. In order to test the QSAR applicability of the present approach, we will develop quantitative models to describe protein stability effects for a complete set of alanine substitutions in the Arc repressor.

2. Theoretical approach

2.1. Protein linear indices of the 'macromolecular pseudograph α-carbon atom adjacency matrix'

The general principles of the linear indices of the 'molecular pseudograph's atom adjacent matrix' for small-to-medium sized organic compounds have been explained in some detail elsewhere. However, an extended overview of this approach will be given in this work.

First, in analogy to the molecular vector X used to represent organic molecules (see Refs. 28 and 29) we introduce here the macromolecular vector (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,[11] side-chain isotropic surface area (ISA), and atomic charges (ECI) of the amino acid,[14] hydropathy index (Kyte–Doolittle scale; HPI),[15] and other hydrophobicity scales such as Hopp and Woods,[16] 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.[11,14] Table 1 depicts several side-chain descriptors for the natural amino acids.[11,14,15]

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

This approach allows us to encode peptides such as VALVGLFVL through out the macromolecular vector $X_m = [-2.69\ 0.07\ -4.19\ -2.69\ 2.23\ -4.19\ -4.92\ -2.69\ -4.19]$, in the z_1-scale (see Table 1). This vector belongs to the product space $\Re^9$. The use of other scales defines alternative macromolecular vectors.

2.2. Local (amino acid) linear indices of the 'macromolecular pseudograph α-carbon atom adjacency matrix'

If a protein consists of n amino acids (*vector of* $\Re^n$), then the k^{th} amino acid linear indices, $\boldsymbol{f}_k(x_{mi})$ are calculated as a linear map on $\Re^n[\boldsymbol{f}_k(x_{mi}) : \Re^n \rightarrow \Re^n$; thus $\boldsymbol{f}_k(x_{mi})$: End on $\Re^n]$ in canonical basis as shown in Eq. 1,

$$\boldsymbol{f}_k(x_{mi}) = \sum_{j=1}^{n} {}^{k}a_{ij}\,{}^{m}X_j \tag{1}$$

where, ${}^{k}a_{ij} = {}^{k}a_{ji}$ (symmetric square matrix), n is the number of amino acids of the protein (α-carbon atom in the protein's backbone), and ${}^{m}X_j$ are the coordinates of the macromolecular vector (X_m) in a system of basis vectors of $\Re^n$. The coordinates of the same vector will be different according to the basis vectors chosen.[34–37] The values of the coordinates thus depend in an essential way on the choice of the basis. With the so-called canonical ('natural') base, e_j denote the n-tuple having 1 in the j^{th} position and 0's elsewhere. In the canonical basis, the coordinates of any vector X coincide with the components of this vector.[34–37] For that reason, those coordinates can be considered as weights (amino-acid labels) of the vertices (α-carbon atoms) of the pseudograph of the protein's backbone.

The coefficients ${}^{k}a_{ij}$ are the elements of the k^{th} power of the macromolecular matrix $\mathbf{M}(G_m)$ of the protein's pseudograph (G_m). Here, $\mathbf{M}(G_m) = [a_{ij}]$, denote the matrix of $\boldsymbol{f}_k(x_{mi})$ with respect to the natural basis. In this matrix n is the number of vertices (α-carbon atoms) of G_m and the elements a_{ij} are defined as follows:

$$\begin{aligned} a_{ij} &= 1 \quad \text{if} \quad i \neq j \quad \text{and} \quad e_k \in E(G_m) \\ &= 1 \quad \text{if} \quad i = j \quad \text{and the amino acid } i \text{ has a hydrogen bond between its side-chain and its main-chain atom} \\ &= 0 \quad \text{otherwise} \end{aligned} \tag{2}$$

where, $E(G_m)$ represents the set of edges of G_m. In this adjacency matrix $\mathbf{M}(G_m)$ the row i and column i correspond to vertex v_i from G_m. The elements $a_{ii} = 1$ are loops in v_i. On the other hand, the element a_{ij} of this matrix represents a bond between an α-carbon atom i and

Table 1. Descriptors for the natural amino acids[11,14–16]

Amino acids		z-Scale[11,14]			Hydrophobicity scale (Kyte–Doolittle)[15]	ISA[14]	ECI[14]
		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

other j. Here, we consider only covalent interaction (peptidic bond) and hydrogen-bond interaction (within a chain as well as between chains). As a first approximation, we considered both interactions to be equivalent, taking into account the 'connectivity of the protein'. The matrix $\mathbf{M}^k(G_m)$ provides the number of walks of length k linking the α-carbon atom of the amino acids i and j. Additionally, proteins containing amino acids having hydrogen bonds between its side-chain and its main-chain atom are represented as a pseudograph. Specifically, the Arc repressor presents this kind of interaction for the amino acid E17, where the presence of this intrasubunit hydrogen bond[31] is accounted by means of a loop in its α-carbon atom of the protein's backbone (see below).

Note, that amino acid's linear indices are defined as a linear transformation $\boldsymbol{f}_k(x_{mi})$ on a macromolecular vector space $\Re^n$. This map is a correspondence that assigns to every vector X_m in $\Re^n$ a vector $\boldsymbol{f}(x_m)$ in such a way that:

$$f(\lambda_1 X_1 + \lambda_2 X_2) = \lambda_1 f(X_1) + \lambda_2 f(X_2) \tag{3}$$

for any scalar λ_1, λ_2, and any vector X_1, X_2 in $\Re^n$. The definition Eq. 1 for $\boldsymbol{f}_k(x_{mi})$ may be written as a single matrix equation:

$$\boldsymbol{f}_k(x_{mi}) = \begin{bmatrix} {}^{m}X'_1 \\ \vdots \\ {}^{m}X'_n \end{bmatrix}^k = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & & \vdots \\ a_{n1} & \cdots & a_{nn} \end{bmatrix}^k \begin{bmatrix} {}^{m}X_1 \\ \vdots \\ {}^{m}X_n \end{bmatrix} \tag{4}$$

or in a more compact form,

$$f_k(x_{mi}) = [{}^{m}X']^k = \mathbf{M}^k(G_m)[{}^{m}X] \tag{5}$$

where $[{}^{m}X]$ is a column vector (a $n \times 1$ matrix) of the coordinates of X_m in the canonical base of $\Re^n$ and $\mathbf{M}^k$, the k^{th} power of the matrix $\mathbf{M}(G_m)$ of the molecular pseudograph (map's matrix). Table 2 exemplifies the calculation of $\boldsymbol{f}_k(x_m)$ for Bradykinin-potentiating pentapeptides previously used in QSAR studies.[14]

2.3. Total (whole-molecule) linear indices of the 'macromolecular pseudograph's α-carbon atom adjacency matrix'

Total protein linear indices are *linear functional* (some mathematicians use the term *linear form*, which means the same as linear functional) on $\Re^n$.[27–30] That is, the k^{th} total protein linear indices are linear maps from $\Re^n$ to the scalar $\Re[f_k(x_m) : \Re^n \to \Re]$. The mathematical definition of these molecular descriptors is the following:

$$f_k(x_m) = \sum_{i=1}^{n} f_k(x_{mi}) \tag{6}$$

where n is the number of amino acids and $\boldsymbol{f}_k(x_{mi})$ are the amino acid's linear indices (linear maps) obtained by Eq. 1. Then, a linear form $\boldsymbol{f}_k(x_m)$ can be written in matrix form,

$$f_k(x_m) = [u]^t[{}^{m}X']^k \tag{7}$$

or

$$f_k(x_m) = [u]^t\mathbf{M}^k[{}^{m}X] \tag{8}$$

for all macromolecular vector $X_m \in \Re^n$. $[u]^t$ is a n-dimensional unitary row vector. As can be seen, the k^{th} total linear indices are calculated by summing the local (amino acid) linear indices of all amino acids in the protein.

2.4. Local (amino acid-type) linear indices of the 'macromolecular pseudograph's α-carbon atom adjacency matrix'

In addition to amino acid linear indices computed for each amino acid in the protein, a local-fragment (amino acid-type) formalism can be developed. The k^{th} amino acid-type linear indices of the 'macromolecular pseudograph's α-carbon atom adjacency matrix' are calculated by summing the k^{th} amino acid linear indices of all amino acids of the same amino acid type in the proteins.

Consequently, if a protein is partitioned in Z molecular fragments, the total protein linear indices can be partitioned in Z local protein linear indices $\boldsymbol{f}_{kL}(x_m)$, $L = 1, \ldots, Z$. That is to say, the total protein linear indices of order k can be expressed as the sum of the local protein linear indices of the Z fragments of the same order:

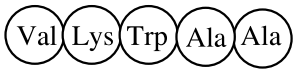
$$f_k(x_m) = \sum_{L=1}^{Z} f_{kL}(x_m) \tag{9}$$

In the amino-acid-type linear indices formalism, each amino acid in the protein is classified into an amino-acid-type (fragment), such as amino acid with R apolar, R polar uncharged, R (+) charged, R (−) charged, 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} fragment (amino-acid-type) linear indices provide much useful information.

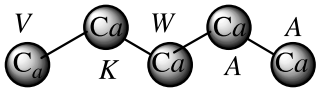
Any local protein linear index has a particular meaning, especially for the first values of k, where the information about the structure of the fragment is contained. Higher values of k relate to the environment information of the fragment considered within the macromolecular pseudograph (G_m).

In any case, whether a complete series of indices is considered, a specific characterization of the chemical structure is obtained (whole protein or fragment), which is not repeated in any other protein. The generalization of the descriptors to 'superior analogs' is necessary for the evaluation of situations where only one descriptor is unable to bring a good structural characterization.[38] The local macromolecular indices can also be used together with the total ones as variables for QSAR/QSPR modeling of properties or activities that depend more on a region or a fragment than on the macromolecule as a whole.

Table 2. Definition and calculation of five ($k = 0-4$) total and local (side-chain amino acid) protein linear indices of the 'macromolecular pseudograph α-carbon atom adjacency matrix' of a Bradykinin-potentiant pentapeptide



Pentapeptide Structure (sequence)



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

Amino-Acid Residue (Side-Chain: R-Group)

Here, we consider only covalent interaction (peptidic bond), but non-covalent interaction (hydrogen-bond and salt bridge interaction) can be taken into consideration (within a chain as well as between chains)

Macromolecular Vector: $\mathbf{X_m} \in \Re^5$

$\mathbf{X_m}$ = [V, K, W, A, A]

In the definition of the $\mathbf{X_m}$, 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, if we write V it means z_1(V), 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^5, the coordinates of any vector $\mathbf{X_m}$ coincide with the components of that macromolecular vector

$[^mX]$ = [-2.69, 2.84, -4.75, 0.07, 0.07]

$[^mX]$: vector of coordinates of $\mathbf{X_m}$ in the Canonical basis of R^5 (a $5x1$matrix)

$$f_0(x_{mi}) = \sum_{j=1}^{n} {}^0a_{ij}\,{}^mX_j = \mathbf{M^0}(\mathrm{Gm})[^mX] = \begin{bmatrix} 1&0&0&0&0\\ 0&1&0&0&0\\ 0&0&1&0&0\\ 0&0&0&1&0\\ 0&0&0&0&1 \end{bmatrix} \begin{bmatrix} V\\ K\\ W\\ A\\ A \end{bmatrix} = \begin{bmatrix} 1V\\ 1K\\ 1W\\ 1A\\ 1A \end{bmatrix}$$

$$f_1(x_{mi}) = \sum_{j=1}^{n} {}^1a_{ij}\,{}^mX_j = \mathbf{M^1}(\mathrm{Gm})[^mX] = \begin{bmatrix} 0&1&0&0&0\\ 1&0&1&0&0\\ 0&1&0&1&0\\ 0&0&1&0&1\\ 0&0&0&1&0 \end{bmatrix} \begin{bmatrix} V\\ K\\ W\\ A\\ A \end{bmatrix} = \begin{bmatrix} 1K\\ 1V+1W\\ 1K+1A\\ 1W+1A\\ 1A \end{bmatrix}$$

$$f_2(x_{mi}) = \sum_{j=1}^{n} {}^2a_{ij}\,{}^mX_j = \mathbf{M^2}(\mathrm{Gm})[^mX] = \begin{bmatrix} 1&0&1&0&0\\ 0&2&0&1&0\\ 1&0&2&0&1\\ 0&1&0&2&0\\ 0&0&1&0&1 \end{bmatrix} \begin{bmatrix} V\\ K\\ W\\ A\\ A \end{bmatrix} = \begin{bmatrix} 1V+1W\\ 2K+1A\\ 1V+2W+1A\\ 1K+2A\\ 1W+1A \end{bmatrix}$$

Amino-acid linear indices of zero, first and second order are a *linear maps*; $f_k(x_{mi})$: $\Re^n \rightarrow \Re^n$ *such that,*

f_0(V, K, W, A, A) = (1V, 1K, 1W, 1A, 1A) = (-2.69, 2.84, -4.75, 0.07, 0.07)

f_1(V, K, W, A, A) = (1K, 1V+1W, 1K+1A, 1W+1A, 1A) = (2.84, -7.44, 2.91, -4.68, 0.07)

f_2(V, K, W, A, A) = (1V+1W, 2K+1A, 1V+2W+1A, 1K+2A, 1W+1A) = (-7.44, 5.75, -12.12, 2.98, -4.68)

and whole-peptide linear indices of zero, first and second order are a *linear functionals*;

$$f_k(x_m) = \sum_{i=1}^{n} f_k(x_{mi})$$

$= f_0(\mathrm{V}) + f_0(\mathrm{K}) + f_0(\mathrm{W}) + f_0(\mathrm{A}) + f_0(\mathrm{A}) = -4.46$

$= f_1(\mathrm{V}) + f_1(\mathrm{K}) + f_1(\mathrm{W}) + f_1(\mathrm{A}) + f_1(\mathrm{A}) = -6.3$

$= f_2(\mathrm{V}) + f_2(\mathrm{K}) + f_2(\mathrm{W}) + f_2(\mathrm{A}) + f_2(\mathrm{A}) = -15.51$

Amino Acid (AA)	${}^{z1}f_{0L}(x_m, AA)$	${}^{z1}f_{1L}(x_m, AA)$	${}^{z1}f_{2L}(x_m, AA)$	${}^{z1}f_{3L}(x_m, AA)$	${}^{z1}f_{4L}(x_m, AA)$
Val (V)	-2.69	2.84	-7.44	5.75	-19.56
Lys (K)	2.84	-7.44	5.75	-19.56	14.48
Trp (W)	-4.75	2.91	-12.12	8.73	-36.36
Ala (A)	0.07	-4.68	2.98	-16.8	11.71
Ala (A)	0.07	0.07	-4.68	2.98	-16.8
Pentapeptide	-4.46	-6.3	-15.51	-18.9	-46.53

3. Results and discussion

3.1. Development of the classification model

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 the 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 (seven having near wild-type stability and five of reduced stability) was used as a test set.

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. Forward stepwise was fixed as the strategy for variable selection. The principle of parsimony (Occams razor) was taken into account as the strategy for model selection. In this connection, we select the functions with the higher statistical significance but having as few parameters (a_k) as possible. The classification model obtained is given below together with the statistical parameters of LDA:

$$\begin{aligned} Class = & -27.661 - 0.308^{Z1} f_0(x_m) \\ & + 0.490^{Z2} f_0(x_m) + 0.219^{\mathbf{HPI}} f_1(x_m) \\ & + 9.304 \times 10^{-11\mathbf{ISA}} f_{15}(x_m) \\ & + 1.272^{\mathbf{ECI}} f_0(x_m) \end{aligned} \quad (10)$$

$$N = 41 \quad \lambda = 0.314 \quad D^2 = 8.72$$
$$F(5, 35) = 15.252 \quad p(F) < 0.0000$$

where λ is Wilks's statistic, D^2 is the squared Mahalanobis distance, and F is the Fisher ratio. 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. These statistics indicate that model (10) is appropriate for the discrimination of near wild-type stability/reduced-stability mutants studied here. The obtained model has a positive predictive value of 95.23% (20/21) of near wild-type stability mutants and a negative predictive value of 100.00% (20/20) of reduced-stability mutants in the training set, for an accuracy (global good classification) of 97.56% (40/41) This model showed a high Matthews' correlation coefficient (MCC) of 0.952; MCC quantified the strength of the linear relation between the molecular descriptors and the classifications. In Table 3 we give the classification of mutants in the training set together with their posterior probabilities calculated from the Mahalanobis distance.

The most important criterion to accept or not a discriminant model, such as model (10), is based on the statistics for the test set. Model (10) classifies correctly 11 of 12 mutants, for an accuracy of 91.67%, with a MCC = 0.837. In Table 3, we give the classification of mutants in the test set. If we considered the data set and the test set (*full* set) the accuracy was 96.23% (51/53).

Canonical analysis is used here to test both the ability of protein linear indices to discriminate between the two

Table 3. Results of the LDA and canonical analyses of the Arc A-mutants in the training and test sets

Mutant	Class[b]	ΔP%[c]	P%(H)[d]	P%(P)[d]	Score[e]
Mutants with near wild-type stability					
1PA8-st6[a]	H	97.98	0.99	0.01	1.64
2SA35-st6	H	99.61	1.00	0.00	2.09
3NA34-st11	H	94.01	0.97	0.03	0.06
4NA11-st6[a]	H	99.20	1.00	0.00	2.49
5QA39-st11	H	33.19	0.67	0.33	−0.17
6GA52-st11	H	85.23	0.93	0.07	−0.23
7KA6-st6[a]	H	60.44	0.80	0.20	0.98
8RA16-st6	H	99.86	1.00	0.00	2.34
9VA25-st6	H	79.15	0.90	0.10	0.92
10MA4-st6	H	61.83	0.81	0.19	0.98
11Arc-st6[a]	H	98.94	0.99	0.01	1.90
12EA27-st6	H	99.70	1.00	0.00	2.43
13KA2-st6	H	98.84	0.99	0.01	2.68
14QA9-st6	H	99.29	1.00	0.00	2.12
15GA3-st6	H	97.39	0.99	0.01	1.98
16MA1-st6[a]	H	61.84	0.81	0.19	0.98
*17Arc-st11	H	−11.43	0.44	0.56	−0.47
18SA5-st6	H	99.86	1.00	0.00	2.32
19RA13-st6	H	99.63	1.00	0.00	2.19
20KA46-st11	H	0.30	0.50	0.50	−0.26
21EA17-st6[a]	H	99.92	1.00	0.00	2.47
22VA18-st6	H	78.84	0.89	0.11	0.92
23RA23-st11	H	74.08	0.87	0.13	−0.01
24KA24-st11	H	79.48	0.90	0.10	0.42
25EA43-st6	H	99.17	1.00	0.00	1.57
26EA28-st11[a]	H	97.49	0.99	0.01	0.19
27MA7-st6	H	60.44	0.80	0.20	0.98
28DA20-st6	H	100.00	1.00	0.00	2.89
Mutants with reduced stability					
29IA51-st11	P	−97.49	0.01	0.99	−1.94
30GA49-st11[a]	P	−60.76	0.20	0.80	−0.16
31LA19-st6	P	−0.18	0.50	0.50	0.48
32GA30-st11	P	−58.82	0.21	0.79	−0.15
33RA50-st11	P	−36.54	0.32	0.68	−0.13
34KA47-st11	P	−1.44	0.49	0.51	−0.27
35PA15-st11[a]	P	−44.82	0.28	0.72	−0.75
36SA44-st11	P	−99.93	0.00	1.00	−2.08
37NA29-st11	P	−71.70	0.14	0.86	−0.25
38VA33-st11	P	−94.26	0.03	0.97	−1.48
39EA48-st11	P	−98.66	0.01	0.99	−1.01
40LA12-st11	P	−99.21	0.00	1.00	−1.90
41FA10-st6[a]	P	−74.79	0.13	0.87	0.33
42LA21-st11	P	−99.16	0.00	1.00	−1.89
43RA31-st11	P	−95.66	0.02	0.98	−0.60
44MA42-st11	P	−98.26	0.01	0.99	−1.50
*45SA32-st1[a]	P	29.74	0.65	0.35	−0.31
46YA38-st11	P	−97.77	0.01	0.99	−1.13
47WA14-st11	P	−99.96	0.00	1.00	−2.45
48RA40-st11	P	−99.17	0.00	1.00	−2.04
49VA22-st11	P	−93.05	0.03	0.97	−1.45
50EA36-st11[a]	P	−12.52	0.44	0.56	−1.16
51IA37-st11	P	−99.59	0.00	1.00	−2.10
52VA41-st11	P	−96.61	0.02	0.98	−1.57
53FA45-st11	P	−99.98	0.00	1.00	−2.30

[*] Mutants that are misclassified by model (10).
[a] Compounds in the test set.
[b] Experimental stability of the Arc A-mutants: H, near wild-type stability mutants; P, reduced-stability mutants.
[c] ΔP% = [P(H-group) − P(P-group)] × 100.
[d] Percentage of probability with which the mutants is predicted as reduced stability/near wild-type stability mutants, respectively.
[e] Canonical scores predicted using canonical analysis (model 11).

groups of the Arc A-mutants and to order these mutants accordingly with their stability profile.

Protein's linear indices and LDA Arc A-Mutant stability canonical analysis principal root:

$$\begin{aligned}\text{Arc mutants} - \text{root} = & -8.636 - 0.155^{Z1} f_0(x_m) \\ & - 0.010^{Z2} f_0(x_m) \\ & + 0.010^{\mathbf{HPI}} f_1(x_m) \\ & + 1.44 \times 10^{-11\mathbf{ISA}} f_{15}(x_m) \\ & + 0.265^{\mathbf{ECI}} f_0(x_m) \qquad (11)\end{aligned}$$

$$N = 41, \quad \lambda = 0.314, \quad R_{canc} = 0.824, \quad \chi^2 = 41.439,$$
$$\text{Mean}(+) = 1.225, \quad \text{Mean}(-) = -1.287,$$
$$p(\chi^2) < 0.0000$$

The canonical transformation of the LDA results yields one canonical root with a good canonical regression coefficient (0.82). Chi-squared test allowed us to test the statistical signification of this analysis with a p-level <0.0000.

When LDA analysis is applied to solve the two-group classification problem, two classification functions are always found.[39,40] Medicinal chemists used to report the function obtained by taking the difference between these two functions when developing QSAR studies.[41–46] However, we cannot use these two classification functions to evaluate all compounds and obtain a bivariate stability map because they are not orthogonal.[40,47] To solve this problem we used canonical analysis. In this case the dimensional reduction caused by canonical analysis makes it possible to obtain a 1-dimension stability map.[47] That is the same that we can order all compounds taking into account its canonical scores. The canonical scores of all A-mutants of the Arc repressor appear in Table 3.

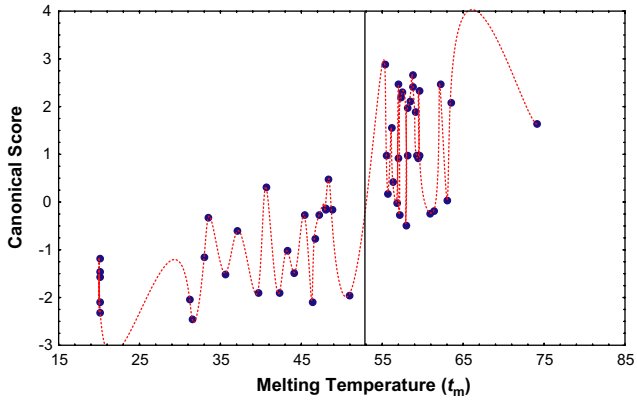


Figure 1. Overall ascendant tendency of canonical scores plotted in the same order in which t_m increases.

We can detect an overall ascendent tendency of canonical scores when they are plotted in the same order as stability (t_m) increases (see Fig. 1). As is expected, the overall mean of canonical root scores for the group of near wild-type stability mutants has an opposite sign (+) with respect to the other group (−).[47]

3.2. Comparison with other approaches

Recently, some in silico techniques have been used to develop classification models that permit us to compute biological stability for each alanine-mutan of the Arc repressor.[48] The relative comparison will be based on the kind of method used for deriving the QSAR and their statistical parameter, the explored molecular descriptors, the overall accuracy (%), Matthews correlation coefficient and the validation method used. Table 4 depicts the comparison between the *TOMOCOMD-CAMPS* method and others reported approaches[48] for the stability of A-mutants of the Arc repressor.

As can be seen, the accuracy in the training set (97.56%) of the *TOMOCOMD-CAMPS* model was higher than

Table 4. Comparison between *TOMOCOMD-CAMPS* method and others approaches for stability of A-mutants of the Arc repressor

Models' features to be compared[a]	Structure-based classification models of the Arc A-mutants stability						
	Linear indices	$\Delta\Theta_0$	D-Fire	Surface	Volume	Log P	Refractivity
Accuracy (%)	97.56	81.1	76.9	70.7	62.3	59.0	60.0
%Nwt[b]	95.23	71.4	92.9	63.6	53.6	80.8	77.3
%RS[b]	100	92.0	58.3	78.9	72.0	15.4	38.9
%NC[b]	0.0	0.0	3.8	22.6	0.0	26.4	24.5
N	41	53	53	53	53	53	53
Wilks' λ(U-statistics)	0.314	0.56	0.79	0.85	0.92	0.99	0.97
F	15.25	39.05	13.9	8.8	4.2	0.5	1.8
p-level	0.00	0.00	0.00	0.00	0.00	0.5	0.2
MCC	0.952	0.643	0.552	0.428	0.259	0.047	0.175
Validation method							
Validation method[c]	i	ii	ii	ii	ii	ii	ii
Accuracy (test set)[d]	91.67	—	—	—	—	—	—
$\%T_{L\text{-}25\%\text{-}O}$[b]	—	79.5	71.8	61.5	56.4	48.7	61.5

[a] Linear 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.[48]

[b] Parameters verifying model quality: %Nwt, %RS, %NC, $\%T_{L\text{-}25\%\text{-}O}$ 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.

of other reported LDA equations (see Table 4). In addition the Wilks' λ statistic for our model was better than those reported in the others models.[48]

Validation of the models is the other major bottleneck in QSAR.[49,50] 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.[49,50] 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.[50] In this case our model shows an accuracy of 91.67% for the test set. It is reasonable to expect some decrease in the 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.

3.3. Modeling the stabilities of a complete set of single alanine-substitution mutants of the Arc repressor of bacteriophage P22

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 the Arc repressor. With this aim, we conform 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 nonaccurate t_m values (<20 °C), which is not useful for MLR analysis.

By using the total protein linear indices of the macromolecular pseudograph's α-carbon atom adjacency matrix and MLR analysis we developed the following QSA(S)R [quantitative structure-activity(stability) relationship] lineal model to describe t_m for these A-mutants of the Arc repressor:

$$\begin{aligned} t_m(^\circ\mathrm{C}) = {} & 31.055(\pm 23.173) \\ & + 3.824(\pm 0.526)^{Z2} f_0(x_m) \\ & + 0.0013(\pm 0.0002)^{\mathbf{ISA}} f_3(x_m) \\ & + 0.192(\pm 0.020)^{\mathbf{HPI}} f_2(x_m) \\ & - 0.929(\pm 0.183)^{Z2} f_1(x_m) \\ & + 2.437(\pm 0.399)^{Z3} f_0(x_m) \\ & - 0.348(\pm 0.060)^{Z3} f_2(x_m) \end{aligned} \quad (12)$$

$N = 45, \quad R = 0.90, \quad R^2 = 0.81, \quad s = 4.29, \quad q^2 = 0.72,$
$s_{cv} = 4.79, \quad F(6.38) = 26.488, \quad p < 0.0000$

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. In Table 5 we give the observed and calculated

Table 5. Experimental and calculated values of melting temperature (t_m) obtained by linear model

Mutant	Obs.[a]	Cal.[b]	Res.[c]	Res.$_{CV}$[d]
1PA8-st6	74.1	*outlier*		
2SA35-st6	63.4	64.8	−1.4	−2.1
3NA34-st11	63.0	58.3	4.7	7.5
4NA11-st6	62.1	54.5	7.6	8.8
5QA39-st11	61.4	60.4	1.0	1.2
6GA52-st11	60.9	63.7	−2.8	−3.8
7KA6-st6	59.6	53.0	6.6	6.9
8RA16-st6	59.5	62.9	−3.4	−4.1
9VA25-st6	59.3	60.3	−1.0	−1.1
10MA4-st6	59.2	52.0	7.2	7.7
11Arc-st6	59.0	59.3	−0.3	−0.3
12EA27-st6	58.8	62.3	−3.5	−3.8
13KA2-st6	58.7	56.7	2.0	2.6
14QA9-st6	58.4	62.1	−3.7	−3.9
15GA3-st6	58.1	62.3	−4.2	−4.6
16MA1-st6	58.0	52.7	5.3	5.6
17Arc-st11	57.9	51.3	6.6	7.3
18SA5-st6	57.5	61.3	−3.8	−4.0
19RA13-st6	57.3	59.0	−1.7	−2.1
20KA46-st11	57.1	*outlier*		
21EA17-st6	57.0	61.0	−4.0	−4.3
22VA18-st6	56.9	57.7	−0.8	−0.8
23RA23-st11	56.7	47.7	9.0	10.6
24KA24-st11	56.3	53.2	3.1	3.4
25EA43-st6	56.1	53.6	2.5	3.0
26EA28-st11	55.7	56.1	−0.4	−0.4
27MA7-st6	55.5	53.0	2.5	2.6
28DA20-st6	55.3	54.4	0.9	1.2
29IA51-st11	50.9	51.9	−1.0	−1.1
30GA49-st11	48.7	51.8	−3.1	−3.4
31LA19-st6	48.3	49.1	−0.8	−0.9
32GA30-st11	47.9	41.4	6.5	8.1
33RA50-st11	47.9	47.6	0.3	0.4
34KA47-st11	47.2	46.3	0.9	1.0
35PA15-st11	46.6	44.6	2.0	2.4
36SA44-st11	46.3	42.3	4.0	6.5
37NA29-st11	45.3	46.6	−1.3	−1.5
38VA33-st11	44.1	45.2	−1.1	−1.3
39EA48-st11	43.2	48.9	−5.7	−6.2
40LA12-st11	42.3	43.1	−0.8	−0.8
41FA10-st6	40.6	42.5	−1.9	−2.3
42LA21-st11	39.6	41.1	−1.5	−1.6
43RA31-st11	37.1	42.8	−5.7	−7.0
44MA42-st11	35.6	42.0	−6.4	−7.0
45SA32-st11	33.5	*outlier*		
46YA38-st11	33.0	37.6	−4.6	−5.5
47WA14-st11	31.5	37.1	−5.6	−8.7
48RA40-st11	31.2	33.6	−2.4	−4.8

[a] Experimental melting temperature, t_m,°C.[54] 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 °C), which is not useful for MLR analysis. st6 and st11 refer to C-terminal sequences of the mutant proteins.[54]

[b] Calculated t_m values by the linear model Eq. 12.

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

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

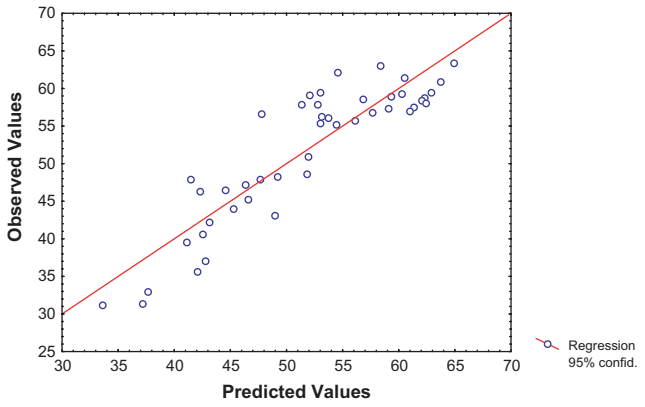


Figure 2. Correlation between experimental and calculated (by Eq. 12) t_m for A-mutants of the Arc repressor.

t_m values by model (12) for the training set, and in Figure 2 is illustrated the linear relationships between them.

Model (12) explains 81% of the variance of the experimental t_m. The predictive ability of model (12) is evidenced by the value of the LOO press statistics (for example $q^2 > 0.5$ and s_{cv}).[49,50]

In developing this model only three mutants (1PA8-st6; 20KA46-st11 and 45SA32-st11) were detected as statistical outliers.[51,52] Outlier detection was carried out using the following standard statistical test: residual, standardized residual, studentized residual, and Cooks' distance.[52] Mutant (PA8) is only significantly more stable than wild type. The t_m of this mutant protein is about 15 °C higher than that of the wild-type parent (see Tables 4 and 5), and the free energy of unfolding is increased by 2.9 kcal mol^{-1} compared with wild type.[53]

Different protein folding may be the reason for the lack of linear regression between protein linear indices and stability (t_m) for these mutants; leading to a nonlinear dependence between t_m and protein linear 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.[53] 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.[53] 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.[53]

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 piecewise method,[39] which produces two linear equations by clustering observations into two groups according to their absolute magnitude. The best fitted piecewise model was

$$\begin{aligned} t_m(^\circ C)_{<BKPT} &= 51.141 + 0.641^{Z2} f_0(x_m) - 0.117^{Z2} f_1(x_m) \\ &\quad + 0.455^{Z3} f_0(x_m) - 0.101^{Z3} f_2(x_m) \\ &\quad + 6.57 \times 10^{-5\ \mathbf{ISA}} f_3(x_m) + 0.03^{\mathbf{HPI}} f_2(x_m) \\ t_m(^\circ C)_{>BKPT} &= 58.741 + 2.201^{Z2} f_0(x_m) - 0.075^{Z2} f_1(x_m) \\ &\quad + 2.459^{Z3} f_0(x_m) - 0.385^{Z3} f_2(x_m) \\ &\quad + 0.000597^{\mathbf{ISA}} f_3(x_m) + 0.184^{\mathbf{HPI}} f_2(x_m) \end{aligned} \quad (13)$$

$$N = 41 \quad R = 0.97 \quad R^2 = 93.43 \quad \text{Bkpt} = 51.87 \quad p < 0.0000$$

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 piecewise hypothesis, p was checked for an absolute value >0.05. The parameter break point (Bkpt) is the t_m value, which marks 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 Table 6, we depict the observed, calculated, and residual values of t_m for the data set. Figure 3 depicts the linear relationships between observed and calculated t_m values in both groups.

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 (51.87), perfectly agrees with an experimental scale previously proposed.[53] 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 mutans' Arc homodimers.

As can be observed in the obtained models, the included variables are related with the factors that influence the stability and this one with the structural features of the Arc dimer. In this sense, the protein's linear indices calculated using z_1, z_2, z_3, ISA, ECI, and HPI values, as amino-acid (side chain) properties are included in most of the developed models. These values are related to hydrophilicity, bulk, and electronic properties. 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.

The preponderance of hydrophobic and electronic effects in the obtained Eqs. (10)–(13) over other types of protein linear indices clearly indicates the importance of the hydrophobic and electronic side-chain factor in the folding of the Arc dimer. This situation means that the stability profile of wild-type Arc and its A-mutants results in topologic/topographic-controlled protein backbone interactions.

Table 6. Experimental and calculated values of melting temperature (t_m) obtained by the nonlinear model

Mutant	Obs.[a]	Cal.[b]	Res.[c]
1PA8-st6	74.1	*outlier*	
2SA35-st6	63.4	60.9	2.5
3NA34-st11	63.0	60.0	3.0
4NA11-st6	62.1	58.0	4.1
5QA39-st11	61.4	58.9	2.5
6GA52-st11	60.9	60.1	0.8
7KA6-st6	59.6	58.1	1.5
8RA16-st6	59.5	57.7	1.8
9VA25-st6	59.3	59.5	−0.2
10MA4-st6	59.2	58.0	1.2
11Arc-st6	59.0	59.3	−0.3
12EA27-st6	58.8	59.2	−0.4
13KA2-st6	58.7	58.7	0.0
14QA9-st6	58.4	59.0	−0.6
15GA3-st6	58.1	59.6	−1.5
16MA1-st6	58.0	58.1	−0.1
17Arc-st11	57.9	58.8	−0.9
18SA5-st6	57.5	59.6	−2.1
19RA13-st6	57.3	57.1	0.2
20KA46-st11	57.1	55.6	1.5
21EA17-st6	57.0	58.9	−1.9
22VA18-st6	56.9	59.0	−2.1
23RA23-st11	56.7	55.9	0.8
24KA24-st11	56.3	57.7	−1.4
25EA43-st6	56.1	56.8	−0.7
26EA28-st11	55.7	58.2	−2.5
27MA7-st6	55.5	58.1	−2.6
28DA20-st6	55.3	57.9	−2.6
29IA51-st11	50.9	49.7	1.2
30GA49-st11	48.7	50.9	−2.2
31LA19-st6	48.3	46.9	1.4
32GA30-st11	47.9	41.7	6.2
33RA50-st11	47.9	47.1	0.8
34KA47-st11	47.2	43.4	3.8
35PA15-st11	46.6	42.8	3.8
36SA44-st11	46.3	45.7	0.6
37NA29-st11	45.3	46.6	−1.3
38VA33-st11	44.1	43.8	0.3
39EA48-st11	43.2	47.1	−3.9
40LA12-st11	42.3	41.7	0.6
41FA10-st6	40.6	39.5	1.1
42LA21-st11	39.6	39.9	−0.3
43RA31-st11	37.1	42.2	−5.1
44MA42-st11	35.6	40.7	−5.1
45SA32-st11	33.5	*outlier*	
46YA38-st11	33.0	35.2	−2.2
47WA14-st11	31.5	32.3	−0.8
48RA40-st11	31.2	30.3	0.9

[a] Experimental melting temperature, t_m (°C).[54] 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 °C), which is not useful for the Piecewise method. st6 and st11 refer to C-terminal sequences of the mutant proteins.[54]

[b] Calculated t_m values by the nonlinear model Eq. 12.

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

4. Concluding remarks

We would expect computational protein science to have a similar effect on the search for new vaccines, receptors, drugs, and so on as molecular modeling and QSAR have had on the search for new drugs. Thus, the definition of novel macromolecular descriptors that could explain different bio-macromolecular properties by means of a QSAR is necessary. In this sense, the approach described here represents a novel and very promising way to bioinformatics research.

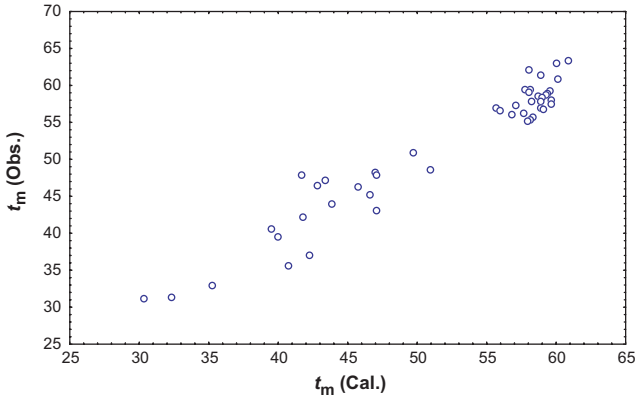


Figure 3. Correlation between experimental and calculated (by Eq. 13) t_m for A-mutants of the Arc repressor.

We have shown here that the use of the protein's total linear 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. These models are not only good enough to predict thermodynamic parameter of the folding of mutants of the Arc dimer repressor, but also permit the interpretation of the driving forces of such folding processes. Nevertheless, future work shall be directed to compare the methodology introduced here with other novel methodologies under the same conditions.

5. Experimental

5.1. *TOMOCOMD-CAMPS* software

TOMOCOMD is an interactive program for molecular design and bioinformatics research.[21] The program is composed of four subprograms, each one of them dealing with drawing structures (drawing mode) and calculating 2D and 3D molecular descriptors (calculation mode). The modules are named CARDD (Computed-Aided 'Rational' Drug Design), CAMPS (Computed-Aided Modeling 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. That is to say, this computer graphics software shows a great efficiency of interaction with the user, without prior knowledge of programming skills (e.g., a practicing pharmaceutic and organic chemist, teacher, university student, and so on). The calculation of total and local macromolecular linear indices for any peptide or protein was implemented in the *TOMOCOMD-CAMPS* software.[21]

5.2. Arc dimer structure and the equilibrium stabilities of a complete set of single alanine-substitution mutants of the Arc repressor of bacteriophage P22

Much work is currently underway to determine the contribution of individual residues to the overall fold and stability of a protein.[54–58] This is a very challenging problem due to the complexity of both the native and unfolded states, and the transition between them. Robert Sauer has done some of the seminal work in this area on the *Arc repressor*.[53,59] This protein provides an attractive system to address this issue because it is small (53 AAs), and amenable to genetic and biophysical studies.[60–62] This is a homodimer protein with a globular domain formed by the intertwining of their monomers. Its secondary structure consists of two anti-parallels β-sheets from residues 8–14, and α-helices formed by residues 15–30 and 32–48.[53]

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. 1b in Ref. 53). Nevertheless, an overview of these electrostatic interactions in the Arc repressor structure will be given. Noncovalent interactions take place:[53]

(i) Between side chains in the same subunit (R16-D20, D20-R23, N29-E36, E36-R31, E36-R40, E43-K46, E43-K47) and; those between side chains in different subunits (E28-R50, R40-S44, R40-F48).
(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).

The data of the Arc repressor mutant was taken from the literature.[53] In this paper, Alanine substitutions were constructed at each of the 51 nonalanine positions in the wild-type Arc sequence. To avoid intracellular proteolysis and purification difficulties,[62,63] these authors constructed the alanine substitution mutant (A-mutants) in backgrounds containing the carboxy-terminal extensions $(His)_6$ (designated st6) or $(His)_6$-Lys-Asn-Gln-His-Glu (designated st11). These tail sequences allow affinity purification, reduce degradation, and cause no significant changes in protein stability.[62]

Milla et al.[53] subjected each purified mutant of the Arc to thermal and urea denaturation experiments. Stability of the proteins was checked by melting temperature (t_m).[53] The values of t_m for 53 Arc homodimers reported by these authors are given in Tables 3–5. The Arc mutants are grouped into two categories (see Table 3): (1) mutants with near wild-type stability and (2) mutants with reduced stability. The first group also includes one mutant with increased stability (PA8-st6). Otherwise, the second one includes five unfolded mutants, even at low temperatures (<20 °C) and absence of denaturants.

In equilibrium and kinetic unfolding–refolding studies only native Arc dimers and denatured monomers are significantly populated. Thus, folding and dimerization are concerted processes.[53,63,64] 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 those 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.[53,55–58]

5.3. Statistical analysis

Linear discrimination analysis (LDA), linear multiple regression (LMR) and the nonlinear estimation analysis, piecewise linear regression (PLR) were used to obtain quantitative models. These statistical analyses were carried out with the STATISTICA software package.[39]

LDA is used in order to generate the classifier function on the basis of the simplicity of the method.[41,65] To test the quality of the derived discriminant functions we used the Wilks' λ and the Mahalanobis distance. The classification of cases was performed by means of the posterior classification probability, which is the probability with which a respective case belongs to a particular group, that is, mutants with near wild-type stability (H) or mutants with reduced stability (P) (see Table 3). In developing this classification function the values of 1 and −1 were assigned to H and P mutants. 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. We also considered the linear discriminant canonical analysis statistics such as the canonical regression coefficient (R_{canc}), chi-squared, and p-level [$p(\chi^2)$].

A simple linear and other more complex nonlinear models were obtained using LMR and PLR as statistic techniques, respectively. The quality of the models was determined by examining the statistic parameters of multivariable comparison of regression and cross-validation procedures. In this sense, the quality of the 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}).[51] In recent years, the LOO press statistics (e.g., q^2) has been used as a means of indicating predictive ability. Many authors consider high q^2 values (for instance, $q^2 > 0.5$) as an indicator or even as the ultimate proof of the high-predictive power of a QSAR model.[49,51]

Acknowledgements

Marrero-Ponce, Y. would like to express his gratitude to Drs. R. Bello and Ricardo Grau, Central University of Las Villas. F. T. acknowledges financial support from the Spanish MCT (Plan Nacional I + D + I, Project No. BQU2001-2935-C02-01).

References and notes

1. Anfinsen, C. B. *Science* **1973**, *181*, 223.
2. Anfinsen, C. B.; Haber, E.; Sela, M.; White, F. H. *Proc. Natl. Acad. Sci. U.S.A.* **1961**, *47*, 1309.
3. Zhang, S. W.; Pan, Q.; Zhang, H. C.; Wu, Y. H.; Shi, J. Y. *Internet Electron. J. Mol. Des.* **2003**, *2*, 392, http://www.biochempress.com.
4. Alberts, B.; Bray, D.; Lewis, J.; Raff, M.; Roberts, K.; Watson, J. D. *Molecular Biology of the Cell*; Garland: New York, 1994.
5. Fersht, A. *Structure and Mechanism in Protein Science: A Guide to Enzyme Catalysis and Protein Folding*; W.H. Freeman: New York, 1999.
6. Freifelder, D. *Molecular Biology. A Comprehensive Introduction to Prokaryotes and Eukaryotes*; Editorial Revolucionaria: Havana, 1983.
7. Lehninger, A. L.; Nelson, D. L.; Cox, M. M. *Principles of Biochemistry*; Worth: New York, 1993.
8. Mathews, C. K.; van Holde, K. E.; Ahern, K. G. *Biochemistry*; Addison Wesley Longan: San Francisco, 2000.
9. Stryer, L. W. H. *Biochemistry*; W.H. Freeman: New York, 1995.
10. Charton, M.; Charton, B. I. *J. Theor. Biol.* **1983**, *102*, 121.
11. Hellberg, S.; Sjöström, M.; Skagerberg, B.; Wold, S. *J. Med. Chem.* **1987**, *30*, 1126.
12. Hellberg, S.; Sjöström, M.; Wold, S. *Acta Chem. Scand.* **1986**, *Sect. B*, 135.
13. Jonsson, J.; Eriksson, L.; Hellberg, S.; Sjöström, M.; Wold, S. *Quant. Struct.-Act. Relat.* **1989**, *8*, 204.
14. Collantes, E. R.; Dunn, W. J., III *J. Med. Chem.* **1995**, *38*, 2705.
15. Kyte, J.; Doolitle, R. F. *J. Mol. Biol.* **1982**, *157*, 105.
16. Hoop, T. P.; Woods, K. R. *Proc. Natl. Acad. Sci. U.S.A.* **1981**, *78*, 3824.
17. Troyer, J. M.; Cohen, E. F. In *Reviews in Computational Chemistry*; Lipkowitz, K. B., Boyd, D. B., Eds.; VHC: New York, 1990; pp 57–80.
18. Leherte, L. *J. Math. Chem.* **2001**, *29*, 47.
19. Cherfils, J.; Duquerroy, S.; Janin, J. *Proteins* **1991**, *11*, 271.
20. Sternberg, M. J. E.; Gabb, H. A.; Jackson, R. *Curr. Opin. Struct. Biol.* **1998**, *8*, 250.
21. Marrero-Ponce, Y.; Romero, V. *TOMOCOMD* software. Central University of Las Villas. **2002**. *TOMOCOMD* (*TO*pological *MO*lecular *COM*puter *D*esign) for Windows, version 1.0 is a preliminary experimental version; in future a professional version will be available upon request from Y. Marrero: yovanimp@qf.uclv.edu.cu; ymarrero77@yahoo.es.
22. Marrero-Ponce, Y. *Molecules* **2003**, *8*, 687, http://www.mdpi.org.
23. Marrero-Ponce, Y. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 2010.
24. Marrero-Ponce, Y.; Castillo-Garit, J. A.; Olazabal, E.; Serrano, H. S.; Morales, A.; Castañedo, N.; Ibarra-Velarde, F.; Huesca-Guillen, A.; Jorge, E.; del Valle, A.; Torrens, F.; Castro, E. A. *J. Comput. Aided Mol. Des.* **2004**, *18*, 615.
25. Marrero-Ponce, Y.; Montero-Torres, A.; Romero-Zaldivar, C.; Iyarreta-Veitía, I.; Mayón Peréz, M.; García-Sánchez, R. *Bioorg. Med. Chem.* Doi: 10.1016/j.bmc.2004.11.008.26.
26. Marrero-Ponce, Y.; Castillo-Garit, J. A.; Torrens, F.; Romero-Zaldivar, V.; Castro, E. *Molecules* **2004**, *9*, 1100.
27. Marrero-Ponce, Y.; Cabrera, M. A.; Romero, V.; Ofori, E.; Montero, L. A. *Int. J. Mol. Sci.* **2003**, *4*, 512.
28. Marrero-Ponce, Y.; Cabrera, M. A.; Romero, V.; González, D. H.; Torrens, F. *J. Pharm. Pharm. Sci.* **2004**, *7*, 186.
29. Marrero-Ponce, Y.; Nodarse, D.; González-Díaz, H.; Ramos de Armas, R.; Romero-Zaldivar, V.; Torrens, F.; Castro, E. *Int. J. Mol. Sci.* **2004**, *5*, 276.
30. Marrero-Ponce, Y.; Castillo-Garit, J. A.; Olazabal, E.; Serrano, H. S.; Morales, A.; Castañedo, N.; Ibarra-Velarde, F.; Huesca-Guillen, A.; Sánchez, A. M.; Torrens, F.; Castro, E. A. *Bioorg. Med. Chem.* **2005**, *13*, 1005.
31. Marrero-Ponce, Y.; Huesca-Guillen, A.; Ibarra-Velarde, F. *J. Mol. Struct. (THEOCHEM)* **2005**, *717*, 67–79.
32. Marrero-Ponce, Y.; González-Díaz, H.; Romero-Zaldivar, V.; Torrens, F.; Castro, E. A. *Bioorg. Med. Chem.* **2004**, *12*, 5331.
33. Marrero-Ponce, Y. *Bioorg. Med. Chem.* **2004**, *12*, 6351.
34. Browder, A. *Mathematical Analysis. An Introduction*; Springer: New York, 1996, pp 176–296.
35. Axler, S. *Linear algebra Done Right*; Springer: New York, 1996; pp 37–70.
36. Ross, K. A.; Wright, C. R. B. *Matemáticas Discretas*; Prentice Hall Hispanoamericana: Mexico DF, 1990.
37. Maltsev, A. I. *Fundamentos del Álgebra Lineal*; Mir: Moscuw, 1976; pp 68–262.
38. Randić, M. *J. Am. Chem. Soc.* **1975**, *69*, 6609.
39. STATISTICA, 1999. *version*. 5.5, Statsoft, Inc.
40. van de Waterbeemd, H. In *Chemometric Methods in Molecular Design*; van de Waterbeemd, H., Ed.; VCH: New York, 1995, pp 265–282.
41. Estrada, E.; Peña, A. *Bioorg. Med. Chem.* **2000**, *8*, 2755.
42. Estrada, E.; Peña, A.; García-Domenech, R. *J. Comput. Aided Mol. Des.* **1998**, *12*, 583.
43. Estrada, E.; Uriarte, E.; Montero, A.; Teijeira, M.; Santana, L.; DeClercq, E. *J. Med. Chem.* **2000**, *43*, 1975.
44. González, D. H.; Marrero-Ponce, Y.; Hernández, I.; Bastida, I.; Tenorio, E.; Nasco, O.; Uriarte, E.; Castañedo, N.; Cabrera, M. A.; Aguila, E.; Marrero, O.; Morales, A.; Pérez, M. *Chem. Res. Toxicol.* **2003**, *16*, 1318.
45. González, H.; Ramos, R.; Molina, R. *Bioinformatics* **2003**, *16*, 2079.
46. Gozalbes, R.; Gálvez, J.; Moreno, A.; Garcia-Domenech, R. *J. Pharm. Pharmacol.* **1999**, *51*, 111.
47. Ford, M. G.; Salt, D. W. In *Chemometric Methods in Molecular Design*; van de Waterbeemd, H., Ed.; Weinheim: New York, 1995; pp 283–292.
48. Ramos de Armas, R.; Gonzalez-Diaz, H.; Molina, R.; Uriarte, E. *Proteins* **2004**, *56*, 715.
49. Golbraikh, A.; Tropsha, A. *J. Mol. Graph. Model.* **2002**, *20*, 269.
50. Wold, S.; Erikson, L. In *Chemometric Methods in Molecular Design*; van de Waterbeemd, H., Ed.; VCH: New York, 1995; pp 309–318.
51. Cronin, M. T. D.; Schultz, T. W. *J. Mol. Struct. (Theochem.)* **2003**, *622*, 39.
52. Belsey, D. A.; Kuh, E.; Welsch, R. E. *Regression Diagnostics*; Wiley: New York, 1980.
53. Milla, M. E.; Brown, M. B.; Sauer, R. T. *Struct. Biol.* **1994**, *1*, 518.
54. Alber, T. *Annu. Rev. Biochem.* **1989**, *58*, 765.
55. Dill, K. A.; Shortle, D. *Annu. Rev. Biochem.* **1991**, *60*, 795.
56. Goldenberg, D. P. *Annu. Rev. Biophys. Biophys. Chem.* **1988**, *17*, 481.
57. Matthews, B. W. *Annu. Rev. Biochem.* **1993**, *62*, 139.
58. Shortle, D. *Curr. Opin. Struct. Biol.* **1993**, *3*, 66.
59. Knight, K. L.; Bowie, J. U.; Vershon, A. K.; Kelley, R. D.; Sauer, R. T. *J. Biol. Chem.* **1989**, *264*, 3639.
60. Bowie, J. U.; Sauer, R. T. *Proc. Natl. Acad. Sci. U.S.A.* **1989**, *86*, 2152.

61. Vershon, A. K.; Bowie, J. U.; Karplus, T. M.; Sauer, R. T. *Proteins* **1986**, *1*, 302.
62. Milla, M. E.; Brown, M. B.; Sauer, R. T. *Protein Sci.* **1993**, *2*, 2198.
63. Bowie, J. U.; Sauer, R. T. *Biochemistry* **1989**, *28*, 7139.
64. Milla, M. E.; Sauer, R. T. *P22. Biochemistry* **1994**, *33*, 1125.
65. McFarland, J. W.; Gans, D. J. In *Comprehesive Medicinal Chemistry*; Hansch, C., Sammes, P. G., Taylor, J. B., Eds.; Pergamon: Oxford, 1990; pp 667–689.