Available online at www.sciencedirect.com
SCIENCE DIRECT®
Bioorganic & Medicinal Chemistry Letters 15 (2005) 5088–5094
Bioorganic & Medicinal Chemistry Letters

Proteins QSAR with Markov average electrostatic potentials

Humberto González-Díaz[a,b,*] and Eugenio Uriarte[a,*]

[a]*Department of Organic Chemistry, Faculty of Pharmacy, University of Santiago de Compostela 15782, Spain*
[b]*Chemical Bioactives Center, Central University of "Las Villas" 54830, Cuba*

Received 10 March 2005; revised 28 June 2005; accepted 5 July 2005
Available online 16 September 2005

Abstract—Classic physicochemical and topological indices have been largely used in small molecules QSAR but less in proteins QSAR. In this study, a Markov model is used to calculate, for the first time, average electrostatic potentials ξ_k for an indirect interaction between aminoacids placed at topologic distances k within a given protein backbone. The short-term average stochastic potential ξ_1 for 53 Arc repressor mutants was used to model the effect of Alanine scanning on thermal stability. The Arc repressor is a model protein of relevance for biochemical studies on bioorganics and medicinal chemistry. A linear discriminant analysis model developed correctly classified 43 out of 53, 81.1% of proteins according to their thermal stability. More specifically, the model classified 20/28, 71.4% of proteins with near wild-type stability and 23/25, 92.0% of proteins with reduced stability. Moreover, predictability in cross-validation procedures was of 81.0%. Expansion of the electrostatic potential in the series ξ_0, ξ_1, ξ_2, and ξ_3, justified the use of the abrupt truncation approach, being the overall accuracy >70.0% for ξ_0 but equal for ξ_1, ξ_2, and ξ_3.The ξ_1 model compared favorably with respect to others based on D-Fire potential, surface area, volume, partition coefficient, and molar refractivity, with less than 77.0% of accuracy [Ramos de Armas, R.; González-Díaz, H.; Molina, R.; Uriarte, E. *Protein Struct. Func. Bioinf.* **2004**, *56*, 715]. The ξ_1 model also has more tractable interpretation than others based on Markovian negentropies and stochastic moments. Finally, the model is notably simpler than the two models based on quadratic and linear indices. Both models, reported by Marrero-Ponce et al., use four-to-five time more descriptors. Introduction of average stochastic potentials may be useful for QSAR applications; having ξ_k amenable physical interpretation and being very effective.

In general, the search novel molecular descriptors to seek quantitative-structure–activity-relationships QSAR[1] nowadays constitutes a widely covered field with more than 1000 molecular descriptors introduced.[2] Nevertheless, 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.[3] More recently, other approaches have been put forward as potential sources for successful biopolymer descriptors, such as Roy et al.,[4] Casanovas et al.,[5] and Leong and Mogenthaler representations;[6] Arteca's average over crossing number,[7,8] Randic's band average widths,[9,10] the sequence-order-coupling numbers,[11,12] α-helix-propensity descriptors, Emini surface index, the SDA sum of cosines of dihedral angles, and Kyle–Doolittle hydrophobicity.[13] One of the most promising applications of QSAR techniques in biochemistry relates to the prediction of protein stability. Proteins must remain stable during biochemistry research and/or biotechnology-related processes.[14]

Numerous researchers worldwide have worked out models to predict the stability of mutants of a wild protein. As shown in Zhou and Zhou's excellent work, a total of 35 proteins with their respective 1023 mutants have been studied, which include all of the examples outlined above.[15–27]

A great deal of work is currently underway to determine the contribution of individual residues to the overall fold and stability of a protein.[28,29] Particularly, great attention has been focused on the *Arc repressors*. This protein provides an attractive system with which to address this issue because it is a small 53 aa and is amenable to biochemistry and biophysical studies. The system is a homodimer protein with a globular domain formed by the intertwining of their monomers. The secondary structure consists of two anti-parallel β-sheets from residues 8 to 14 and α-helices formed by residues 15–30 and 32–48.[30]

Keywords: MARCH-INSIDE; Protein stability; QSAR; Electrostatic potential; Markov model.
* Corresponding authors. Tel.: +34 98 15 63 10 0x14 93 8; fax: +34 9 81 59 49 12; e-mail addresses: gonzalezdiazh@yahoo.es; qofuri@usc.es

0960-894X/$ - see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.bmcl.2005.07.056

Nevertheless, neither Zhou and Zhou's work nor, other previous studies any reported in the literature have attempted to predict the stability of Arc repressors before 2003.[15–27] Until our concern, Ramos de Armas et al. reported for the first time a QSAR predicting stability for Arc mutants.[31] The model is relatively simple in statistical terms, with only one variable, but the molecular descriptor has to be interpreted in terms of Shannon entropy in such a way that it is not very amenable. A second model reported to predict the thermal stability of Arc mutants has been just introduced by Gonzalez-Díaz et al.[32] The model is also simple but does not process a very sound physical interpretation based on stochastic spectral moments. In the present work, we have addressed this aspect and other issues such as introducing a new Markov model with high accuracy but having direct physical interpretation in terms of electrostatic potential.

This approach used a Markov chain (MC) model[33] to codify information about the proteins' molecular structure and constitutes one generalization of other molecular descriptors derived with the so-called MARCH-INSIDE MARkovian CHemicals IN SIlico DEsign approach.[34] A precise definition of different molecular descriptors generated by this methodology can be found in the literature.[35,36]

Herein, the method uses as a source of molecular descriptor the ${}^{1}\Pi$ matrix the short-term electrostatic interaction matrix built up as a squared matrix $n \times n$, being n the number of aminoacids aa in the protein. One can consider a hypothetical situation in which every jth-aa has an electrostatic potential ϕ_j at an arbitrary initial time t_0. All these potentials can be listed as elements of the vector ${}^{0}\phi$. It can be supposed that, after this initial situation, all the amino acids interact with electrostatic energy ${}^{1}E_{ij}$ with every other aa_j in the protein. For the sake of simplicity, a truncation function α_{ij} is applied in such a way that short-term electrostatic interaction takes place only between neighboring amino acids $\alpha_{ij} = 1$. Otherwise, the electrostatic interaction banished $\alpha_{ij} = 0$.[37]

Ignoring direct interaction between distant amino acids does not prevent any that electrostatic interactions from propagating between those amino acids within the protein backbone in an indirect manner. Thus, by using the MC theory it is possible to develop a simpler model to calculate the average electrostatic potentials ξ_k for the indirect interaction between any aa_j and the others aa_i placed at a distance k within the protein backbone:[38]

$$\xi_k = \sum_{j=1}^{n} {}^{A}p_k(j) \cdot \varphi_j = {}^{0}\pi^{T} \cdot {}^{k}\Pi \cdot {}^{0}\varphi = {}^{0}\pi^{T} \cdot ({}^{1}\Pi)^{k} \cdot {}^{0}\varphi \quad (1)$$

It is remarkable that the average electrostatic potentials ξ_k depend on the absolute probabilities ${}^{A}p_k(j)$ with which the amino acids interact with other amino acids placed at distance k. The potential ξ_k also depends on the initially unperturbed electrostatic potential of the aminoacid. In matrix form represented above, the ${}^{A}p_k(j)$ are calculated with the vector ${}^{0}\pi$, of absolute initial probabilities, and the matrix ${}^{1}\Pi$ using the Chapman–Kolgomorov equations.[39] In particular, the evaluation of such expansions for $k = 0$ gives the initial average unperturbed electrostatic potential ξ_0, for $k = 1$ the short-range potential ξ_1, for $k = 2$ the middle-range potential ξ_2, and for $k = 3$ the long-range one. We illustrate, this expansion for the tripeptide Ala-Val-Trp (AVW) as follows:

$$\xi_0 = \left[{}^{A}p_0(\mathrm{A}), {}^{A}p_0(\mathrm{V}), {}^{A}p_0(\mathrm{W})\right] \cdot \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} \varphi_A \\ \varphi_V \\ \varphi_W \end{bmatrix} = {}^{A}p_0(\mathrm{A}) \cdot \varphi_A + {}^{A}p_0(\mathrm{V}) \cdot \varphi_V + {}^{A}p_0(\mathrm{W}) \cdot \varphi_W \quad (2a)$$

$$\xi_1 = \left[{}^{A}p_0(\mathrm{A}), {}^{A}p_0(\mathrm{V}), {}^{A}p_0(\mathrm{W})\right] \cdot \begin{bmatrix} {}^{1}p_{\mathrm{AA}} & {}^{1}p_{\mathrm{AV}} & 0 \\ {}^{1}p_{\mathrm{VA}} & {}^{1}p_{\mathrm{VV}} & {}^{1}p_{\mathrm{VW}} \\ 0 & {}^{1}p_{WV} & {}^{1}p_{WW} \end{bmatrix} \times \begin{bmatrix} \varphi_A \\ \varphi_V \\ \varphi_W \end{bmatrix} \quad (2b)$$

$$\xi_2 = \left[{}^{A}p_0(\mathrm{A}), {}^{A}p_0(\mathrm{V}), {}^{A}p_0(\mathrm{W})\right] \cdot \begin{bmatrix} {}^{1}p_{\mathrm{AA}} & {}^{1}p_{\mathrm{AV}} & 0 \\ {}^{1}p_{\mathrm{VA}} & {}^{1}p_{\mathrm{VV}} & {}^{1}p_{\mathrm{VW}} \\ 0 & {}^{1}p_{WV} & {}^{1}p_{\mathrm{WW}} \end{bmatrix} \times \begin{bmatrix} {}^{1}p_{\mathrm{AA}} & {}^{1}p_{\mathrm{AV}} & 0 \\ {}^{1}p_{\mathrm{VA}} & {}^{1}p_{\mathrm{VV}} & {}^{1}p_{\mathrm{VW}} \\ 0 & {}^{1}p_{\mathrm{WV}} & {}^{1}p_{\mathrm{WW}} \end{bmatrix} \cdot \begin{bmatrix} \varphi_A \\ \varphi_V \\ \varphi_W \end{bmatrix} \quad (2c)$$

$$\xi_3 = \left[{}^{A}p_0(\mathrm{A}), {}^{A}p_0(\mathrm{V}), {}^{A}p_0(\mathrm{W})\right] \cdot \begin{bmatrix} {}^{1}p_{\mathrm{AA}} & {}^{1}p_{\mathrm{AV}} & 0 \\ {}^{1}p_{\mathrm{VA}} & {}^{1}p_{\mathrm{VV}} & {}^{1}p_{\mathrm{VW}} \\ 0 & {}^{1}p_{WV} & {}^{1}p_{\mathrm{WW}} \end{bmatrix} \times \begin{bmatrix} {}^{1}p_{\mathrm{AA}} & {}^{1}p_{\mathrm{AV}} & 0 \\ {}^{1}p_{\mathrm{VA}} & {}^{1}p_{\mathrm{VV}} & {}^{1}p_{\mathrm{VW}} \\ 0 & {}^{1}p_{\mathrm{WV}} & {}^{1}p_{\mathrm{WW}} \end{bmatrix} \cdot \begin{bmatrix} {}^{1}p_{\mathrm{AA}} & {}^{1}p_{\mathrm{AV}} & 0 \\ {}^{1}p_{\mathrm{VA}} & {}^{1}p_{\mathrm{VV}} & {}^{1}p_{\mathrm{VW}} \\ 0 & {}^{1}p_{WV} & {}^{1}p_{WW} \end{bmatrix} \times \begin{bmatrix} \varphi_A \\ \varphi_V \\ \varphi_W \end{bmatrix} \quad (2d)$$

To carry out the calculations referred to in Eq. 1 and detailed in Eqs. 2a–2d, and the elements ${}^{1}p_{ij}$ of ${}^{1}\Pi$ and the absolute initial probabilities ${}^{A}p_k(j)$ were calculated as: [40,41]

$${}^{1}p_{ij} = \frac{\alpha_{ij} \cdot E_{ij}}{\sum_{m=1}^{\delta+1} \alpha_{im} \cdot E_{ij}} = \frac{\alpha_{ij} \cdot \frac{q_i \cdot q_j}{d_{ij}^2}}{\sum_{m=1}^{\delta+1} \alpha_{im} \cdot \frac{q_i \cdot q_m}{d_{ij}^2}} = \frac{\alpha_{ij} \cdot q_i \cdot \frac{q_j}{d_{ij}^2}}{q_i \cdot \sum_{m=1}^{\delta+1} \alpha_{im} \cdot \frac{q_m}{d_{ij}^2}} = \frac{\alpha_{ij} \cdot q_j}{\sum_{m=1}^{\delta+1} \alpha_{im} \cdot q_m} \quad (3)$$

$${}^{A}p_0(j) = \frac{q_j}{\sum_{m=1}^{n} q_m} \quad (4)$$

where, for the *i*th-aa and the *j*th-aa: the neighborhood relationship truncation function $\alpha_{ij} = 1$ turned if there were a peptidic or hydrogen bond, $d_{ij} = 1$ is the topologic distance, and q_i and q_j are the electronic charges.[42] All calculations of molecular indices for Arc mutants were carried out with our software BIOMARKS 1.0® (BIOinformatics MARKovian Studio), see Figure 1.[43]

Electrostatics is the theory of a static configuration of charges. Protein electrostatics is a very covered field nowadays with incidence in different aspects of biochemistry.[44] For instance, the nature of electrostatic barrier for proton transport in aquaporins has been analyzed by semimacroscopic and microscopic models recently.[45] In this sense, the truncation of the electrostatic field is usually applied to simplify all the calculations in large biologic systems as proteins are.[46–49] The review after Norbeg and Nilsson constitutes a seminar work on the topic.[50] Truncation is usually applied in molecular dynamic studies together with spectroscopy for protein structure characterization as in the works by Celda's group.[51–53]

On the other hand, MC models are well-known tools for analyzing biological sequence data.[54,55] Another use of these models is data-based searching and multiple sequence alignment of protein families.[56] Protein-turn types,[57] sub-cellular locations,[58,59] and secondary structure[60] have been successfully predicted. Krogh et al.[61] have also proposed a hidden Markov Model architecture in bioinformatics. In addition, Markov's stochastic process has been used for protein folding recognition[62] and prediction of protein signal sequences.[63,64] Seminar works after Choucan be found to be related to the application of MC theory in biochemistry and bioinformatics.[65–68]

In this work, we used the MC model to derive average electrostatic potentials considering non-interacting aminoacids ξ_0, short-range ξ_1, middle-range ξ_2, and long-range electrostatic interactions ξ_3 for 53 Arc repressor mutants. All the 53 mutants and its classification within near wild type (nwt) or decreased stability (ds) group were taken from the literature.[30–32] These aforementioned descriptors were used to carry out a linear discriminant analysis (LDA) analysis to classify each mutant with respect to its thermal stability and the best model found was:

$$\text{Stability} = -1.86 \times \xi_1 + 0.06 \tag{5}$$

$$N = 53 \quad \lambda = 0.63 \quad F(2, 50) = 29.57 \quad p < 0.001,$$

$$\% = 81.1 \quad \%^{+} = 71.4 \quad \%^{-} = 92.0,$$

where N is the number of proteins used in the study including alanine-mutants and the wild-type Arc (wtArc) repressor. The statistical parameters of the above equation were also shown including Wilk's statistic λ, Fischer ratio F, and significance level p.[69] The discriminant function classified correctly 43 out of 53 mutant proteins according to their relative stability related to wild-type protein. This provides a level of accuracy of 81.1%. More specifically, the model classified 20/28 proteins' nwt stability, 71.4%$^{+}$ and 23/25, 92.0%$^{-}$ proteins within ds group. Table 1 shows the respective classification matrices for training, as well as cross-validation.

A cross-validation procedure was subsequently performed for assess model predictability. This cross-validation was carried out by using the average of a resubstitution technique composed by some main stages. First, we single out at random 25% of the compounds and constitute the first predicting series cv1. Afterwards, compounds in predicting series are interactively interchanged with those in training ones, creating three additional predicting series cv2, cv3, and cv4. Finally,

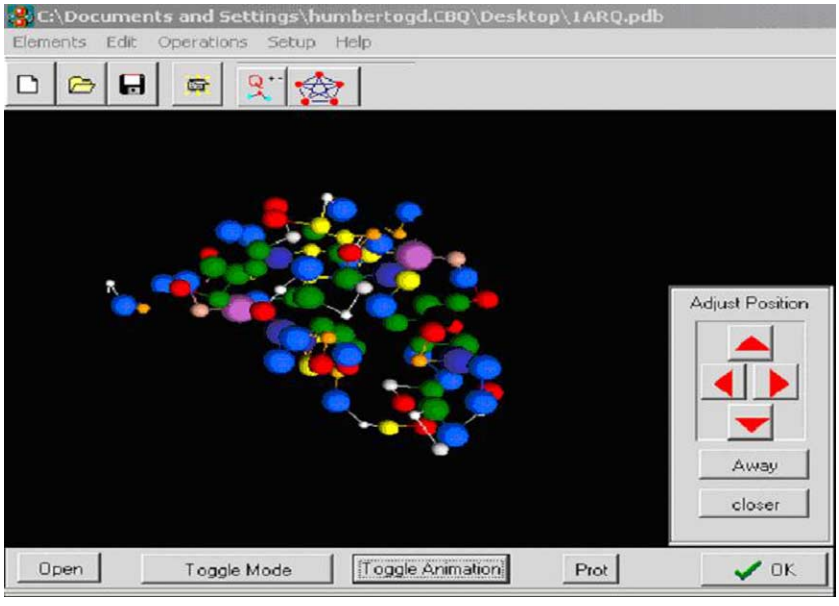

Figure 1. Arc wild type protein depicted at BIOMARKS interface.

Table 1. Accuracy for training set and resubstitution cross-validation using ξ_1

	%	nwt	ds
Train			
nwt	71.4	**20**	8
ds	92	2	**23**
Total	81.1		
cv-average			
nwt	71.2	**15**	6
ds	92.1	2	**17**
Total	81.0		
cv1			
nwt	68.2	**15**	7
ds	88.9	2	**16**
Total	77.5		
cv2			
nwt	73.7	**14**	5
ds	90.0	2	**18**
Total	82.1		
cv3			
nwt	66.7	14	7
ds	94.7	1	18
Total	80.0		
cv4			
nwt	77.3	**17**	5
ds	94.4	1	**17**
Total	85.0		

nwt: near-wild-type proteins, ds: decreased stability proteins. %: accuracy.

we reported the accuracy, and classification matrices for each series, and averaged results. The present model has shown quite a good average predictability (cv-average) of 81.0 %. In particular, the model showed a very high average accuracy of 92.1%, predicting the stability class of ds mutants. Near-wild-type mutants are predicted with a slightly lower average accuracy 71.2%, which is, however, a significant result. The present level of accuracy has usually been considered as very good for researchers using molecular descriptors used in LDA in QSAR studies, as were the cases of Cabrera-Pérez et al.,[70,71] Molina et al.,[72] and González et al.[73] The importance of this result also relates to the simplicity of the present topologic methodology, which does not need any 3D exhaustive structural information. The present result coincides with that reported by Ramos de Armas et al. on the use of MC models to encode proteins and peptide structure in QSAR studies.[74,75] Tables 1 and 2, presented as supplementary materials, show detailed results for each protein, in connection with the observed classification versus predicted training, and cross-validation probabilities. Figure 2 graphically illustrates the average overall predictability results.

Finally, the physical interpretation of the present MC model may be of major interest. First, a direct inspection of Eq. 5 shows a unitary increase in ξ_1 decreased 1.86 times negative coefficient the possibilities of a protein to remain stable. This fact can be explained taking into consideration that those protein mutants with very high electrostatic interactions more easily lose stability.[76] Second, we derived equations similar to 5 for ξ_0, ξ_2, and ξ_3. The model considered does not interact with amino acids ξ_0 presenting less than 70% of overall accuracy, indicating that electrostatic interactions, in fact, play an important role Arc repressors in stability. On the other hand, the results for middle-range ξ_2 and long-range potentials ξ_3 justified the use of the abrupt truncation function α_{ij}[77] in the study of Arc repressors, given that no additional improvement of the model was found. The present model diverges in physical terms than others reported by Ramos de Armas et al. that use stochastic entropies $\Delta\Theta_0$, ignoring at all the possibility of interaction between aminoacids.[31] However, this

Table 2. Comparative study with other nine stability scoring functions

Stat[a]	Physicochemical parameters					Algebraic forms		Markov indices		
	DF[b]	SA[c]	V[d]	$\log P$[e]	M_R[f]	q[g]	f[g]	$\Delta\Theta_0$[h]	$^{SR}\pi_1$[i]	ξ_1
Nmd	1	1	1	1	1	4	5	1	1	1
RP	−5.5	−14.7	−30.2	−37.5	−35.2	5.0	16.9	0.0	−2.4	0.0
%T	76.9	70.7	62.3	59.0	60.0	85.4	97.6	81.1	79.2	81.1
%nwt	92.9	63.6	53.6	80.8	77.3	85.0	95.2	71.4	67.8	71.4
%RS	58.3	78.9	72.0	15.4	38.9	85.7	100	92.0	92.0	92.0
%T_{cv}	71.8	61.5	56.4	48.7	61.5	80.5	91.7	79.5	79.2	81.1
p	0.001	0.001	0.001	0.5	0.2	0.01	0.01	0.0	0.0	0.0

[a] Statistical parameters verifying model quality are: the number of molecular descriptors in the model (nmd), total (%T), near wild-type group (%nwt), reduced stability group (%DS), and average cross-validation (%T_{cv}) percentages of good classification.
[b] D-Fire potential DF.
[c] Surface area SA.
[d] Volume V.
[e] Logarithm of the partition coefficient $\log P$.
[f] Molar refractivity M_R.
[g] These very complicated models are based on four-to-five times more molecular descriptors named: quadratic indices ($^{Z3}q_0(\chi_m)$, $^{Z2}q_7(\chi_m)$, $^{Z1}q_1(\chi_m)$, and $^{Z2}q_2(\chi_m)$), for the first model, and linear indices ($^{Z1}f_0(\chi_m)$, $^{Z2}f_0(\chi_m)$, $^{HPI}f_1(\chi_m)$, $^{HSA}f_{15}(\chi_m)$, and $_{EC1}f_0(\chi_m)$), for the second one, which lack any direct physical sense.
[h] Markovian negentropies.
[i] Stochastic moments.

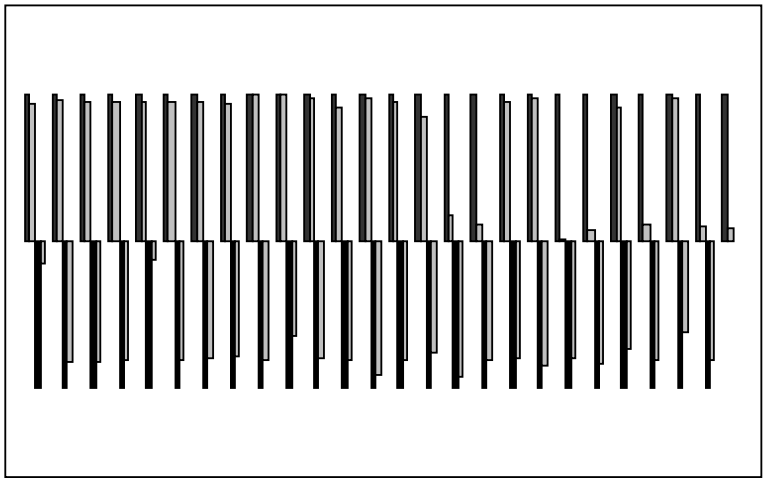

Figure 2. Overall average predictability of the model in terms of predicted probability (y) for each kind of mutation versus protein backbone position (x). Upper half shows nwt group and bottom half ds group. Observed probabilities (in black) are shown for the purpose of comparison being set equal to 100% (certainty of occurrence) predicted probabilities are depicted in gray.

model coincides with the others reported by González-Díaz et al.,[32] which have taken into consideration short-range interaction moments ${}^{SR}\pi_1$, see Table 2.

In closing, we carried out a comparison of the present model with other models predicting the stability of Arc repressor mutants. As shown in Table 2 in general one-descriptor models based on classic physicochemical and geometric parameters, such as surface area SA, volume V, partition coefficient $\log P$, molar refractivity M_R, and D-Fire potential DF, presented a weak linear relationship with Arc repressor stability compared to MC models. The parameters DF, SA, and V have shown less than 77% percentages of good classification compared with more than 80% of the MC models. On the other hand, $\log P$ and M_R didnot present any significant relationship with Arc repressor stability $p > 0.05$.[31,32,78]

Finally, the present model very favorably compares in terms of simplicity with other two models that have been very recently reported by Marrero-Ponce et al.[79,80] These very complicated models are based on four quadratic indices (${}^{Z3}q_0(\chi_m)$, ${}^{Z2}q_7(\chi_m)$, ${}^{Z1}q_1(\chi_m)$, and ${}^{Z2}q_2(\chi_m)$) for the first model,[79] and five linear indices (${}^{Z1}f_0(\chi_m)$, ${}^{Z2}f_0(\chi_m)$, ${}^{HPI}f_1(\chi_m)$, ${}^{HSA}f_{15}(\chi_m)$, and ${}^{ECI}f_0(\chi_m)$) for the second one.[80] In Table 2, one can note that all the one-variable models have a much more decreased relative predictability (RP) than our model. Conversely, the models reported by Marrero-Ponce et al. using five-to-four time more molecular descriptors than the one-variable (ξ_1) model increase the overall predictability by only 5.0% or 16.9%, respectively. The RP values were determined with respect to the model's overall predictability as $RP = (\%T - 81.1)*100/\%T$ (see Table 2). These results indicate that these models seem to be over-fitted with respect to the others due to the very large number of parameters used versus a 'poor' improvement of accuracy.

As a sort of concluding remarks in first instance, the present work introduces, for the first time, a method to derive average electrostatic potentials ξ_k with MC models for proteins QSAR in bioorganic chemistry. The paper also introduces a novel method to classify Arc repressor mutants with respect to their stability. The present model in a more precise physical theoretic context gives a higher importance to electrostatic interactions for the stability of Arc repressor than that reported by Ramos de Armas.[31] However, it confirms the necessity of truncation approaches dispensing with the calculation of long-range electrostatic interactions.[49–53] This result coincides in spirit with those of Gonzalez and Morales et al.,[81–83] on the application of QSAR to problems at the border line between bioorganic chemistry and polymer sciences. After a visual inspection of ξ_k equations, one can detect a vector–matrix–vector form that determines certain contact points with classic topologic indices.[84,85] However, this work lays specific emphasis on expanding the possibilities, outlined before, of MC models to derive molecular descriptors[86–88] making use of physicochemical theories, such as thermodynamics and/or electrostatics,[89,90] in this case.[37,38,91]

References and notes

1. Kubinyi, H.; Taylor, J.; Ramdsen, C. Quant. Drug Des. In Hansch, C., Ed.; Compr. Med. Chem.; Pergamon press: New York, 1990; Vol. 4, pp 589–643.
2. Todeschini, R.; Consonni, V. Handbook of Molecular Descriptors; Wiley VCH: Weinheim, Germany, 2000.
3. Flory, P. J. Principles of Polymer Chemistry; Cornell University Press: Itaha, 1953.
4. Roy, A.; Raychaudhury, C.; Nandy, A. *J. Biosci.* **1998**, *23*, 55.
5. Casanovas, J.; Miro-Julia, J.; Rosselló, F. *J. Math. Biol.* **2003**, *47*, 1.
6. Leong, P. M.; Mogenthaler, S. *Comput. Appl. Biosci.* **1995**, *12*, 503.
7. Arteca, G. A. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 550.
8. Arteca, G. A.; Mezey, P. G. *J. Mol. Graphics* **1990**, *8*, 66.
9. Randič, M.; Vračko, M.; Nandy, A.; Basak, S. C. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1235.

10. Randič, M.; Balaban, A. T. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 532.
11. Hua, S.; Sun, Z. *Bioinformatics* **2001**, *17*, 721.
12. Cai, Y-D.; Lina, S. L. *BBA* **2003**, *1648*, 127.
13. Lejon, T.; Strom, B. M.; Svensen, S. J. *J. Pept. Sci.* **2002**, *7*, 74.
14. EUFEPS Announcement. *Eur. J. Pharm. Sci.* **2002**, *15*, 101.
15. Zhou, H.; Zhou, Y. *Proteins: Struc. Funct. Genet.* **2002**, *49*, 483.
16. Green, S. M.; Meeker, A. K.; Shortle, D. *Biochemistry* **1992**, *31*, 5717.
17. O'Neil, K. T.; De Grado, W. F. *Science* **1990**, *250*, 646.
18. Blaber, M.; Zang, X.; Matthews, B. W. *Science* **1993**, *260*, 1637.
19. Kim, D. E.; Fisher, C.; Baker, D. *J. Mol. Biol.* **2000**, *298*, 971.
20. Hamill, S. J.; Steward, A.; Clarke, J. *J. Mol. Biol.* **2000**, *297*, 165.
21. Fulton, K. F.; Main, E. R. G.; Daggett, V.; Jackson, S. E. *J. Mol. Biol.* **1999**, *291*, 445.
22. Kragelund, B. B.; Osmark, P.; Neergaard, T. B.; Schidt, J.; Kristiansen, K.; Knudsen, J.; Poulsen, F. M. *Nat. Struct. Biol.* **1999**, *6*, 594.
23. Ternström, T.; Mayor, U.; Akke, M.; Oliveberg, M. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 14854.
24. Lorch, M.; Mason, J. M.; Clarke, A. R.; Parker, M. J. *Biochemistry* **1999**, *38*, 1377.
25. Julenius, K.; Thulin, E.; Linse, S.; Finn, B. E. *Biochemistry* **1998**, *37*, 8915.
26. Alber, T. A. *Rev. Biochem.* **1989**, *58*, 765.
27. Dill, K. A.; Shortle, D. A. *Rev. Biochem.* **1991**, *60*, 795.
28. Alber, T. *Annu. Rev. Biochem.* **1989**, *58*, 765.
29. Dill, K. A.; Shortle, D. *Annu. Rev. Biochem.* **1991**, *60*, 795.
30. Milla, M. E.; Brown, M. B.; Sauer, R. T. *Struct. Biol.* **1994**, *1*, 518.
31. Ramos de Armas, R.; González-Díaz, H.; Molina, R.; Uriarte, E. *Proteins: Struct. Func. Bioinf.* **2004**, *56*, 715.
32. González-Díaz, H.; Uriarte, E.; Ramos de Armas, R. *Bioorg. Med. Chem.* **2004**, *13*, 323.
33. Freund, J. A.; Poschel, T. Stochastic processes in physics, chemistry, and biology. *Lect. Notes Phys.*; Springer: Berlin, Germany, 2000.
34. González-Díaz, H.; Olazábal, E.; Castañedo, N.; Hernández, S. I.; Morales, A.; Serrano, H. S.; González, J.; Ramos de Armas, R. *J. Mol. Mod.* **2002**, *8*, 237.
35. González-Díaz, H.; Gia, O.; Uriarte, E.; Hernández, I.; Ramos, R.; Chaviano, M.; Seijo, S.; Castillo, J. A.; Morales, L.; Santana, L.; Akpaloo, D.; Molina, E.; Cruz, M.; Torres, L. A.; Cabrera, M. A. *J. Mol. Mod.* **2003**, *9*, 395.
36. González-Díaz, H.; Marrero, Y.; Hernández, I.; Bastida, I.; Tenorio, I.; Nasco, O.; Uriarte, E.; Castañedo, N. C.; Cabrera-Pérez, M. A.; Aguila, E.; Marrero, O.; Morales, A.; González, M. P. *Chem. Res. Toxicol.* **2003**, *16*, 1318.
37. González-Díaz, H.; Molina, R.; Uriarte, E. *Bioorg. Med. Chem. Lett.* **2004**, *14*, 4691.
38. González-Díaz, H.; Molina, R.; Uriarte, E. *Polymer* **2004**, *45*, 3845.
39. González-Díaz, H.; Bastida, I.; Castañedo, N.; Nasco, O.; Olazabal, E.; Morales, A.; Serrano, H. S.; Ramos de Armas, R. *Bull. Math. Biol.* **2004**, *66*, 1285.
40. González-Díaz, H.; Uriarte, E. *Biopolymers* **2005**, *77*, 296.
41. González-Díaz, H.; Ramos de Armas, R.; Molina, R. *Bioinformatics* **2003**, *19*, 2079.
42. Collantes, E. R.; Dunn, W. J. *J. Med. Chem.* **1995**, *38*, 2705.
43. González-Díaz, H.; Hernández, I. *BIOMARKS* **2002**, *version 1.0*, This is a preliminary experimental version. A future professional version shall be available to the public. For any information about it, send and e-mail to the corresponding author gonzalezdiazh@yahoo.es or qohumbe@usc.es.
44. Kundu, S.; Gupta-Bhaya, P. *J. Mol. Struct. (THEOCHEM)* **2004**, *668*, 65.
45. Burykin, A.; Warshel, A. *FEBS Lett.* **2004**, *570*, 41.
46. Sagui, C.; Darden, T. A. *Annu. Rev. Biophys. Biomol. Struct.* **1999**, *28*, 155.
47. Guenot, J.; Kollman, P. A. *J. Comp. Chem.* **1993**, *14*, 295.
48. Harvey, S. C. *Proteins* **1989**, *5*, 78.
49. Auffinger, P.; Beveridge, D. L. *Chem. Phys. Lett.* **1995**, *234*, 413.
50. Norberg, J.; Nilsson, L. *Biophys. J.* **2000**, *79*, 1537.
51. Navarro, E.; Fenude, E.; Celda, B. *Biopolymers* **2004**, *73*, 229.
52. Navarro, E.; Fenude, E.; Celda, B. *Biopolymers* **2002**, *64*, 198.
53. Monleon, D.; Celda, B. *Biopolymers* **2003**, *70*, 212.
54. Vorodovsky, M.; Koonin, E. V.; Rudd, K. E. *Trends Biochem. Sci.* **1994**, *19*, 309.
55. Vorodovsky, M.; Macininch, J. D.; Koonin, E. V.; Rudd, K. E.; Médigue, C.; Danchin, A. *Nucleic Acid Res.* **1995**, *23*, 3554.
56. Krogh, A.; Brown, M.; Mian, I. S.; Sjeander, K.; Haussler, D. *J. Mol. Biol.* **1994**, *235*, 1501.
57. Chou, K-C. *Biopolymers* **1997**, *42*, 837.
58. Yuan, Z. *FEBS Lett.* **1999**, *451*, 23.
59. Hua, S.; Sun, Z. *Bioinformatics* **2001**, *17*, 721.
60. Hubbard, T. J.; Park, J. *Proteins: Struct. Funct. Genet.* **1995**, *23*, 398.
61. Krogh, A.; Brown, M.; Mian, I. S.; Sjeander, K.; Haussler, D. *J. Mol. Biol.* **1994**, *235*, 1501.
62. Di Francesco, V.; Munson, P. J.; Garnier, J. *Bioinformatics* **1999**, *15*, 131.
63. Chou, K-C. *Curr. Protein Pept. Sci.* **2002**, *3*, 615.
64. Chou, K-C. *Peptides* **2001**, *22*, 1973.
65. Chou, K.-C. *Anal. Biochem.* **2002**, *86*, 1.
66. Chou, K.-C. *J. Biol. Chem.* **1993**, *268*, 16938.
67. Chou, K.-C. *Anal. Biochem.* **1996**, *233*, 1.
68. Chou, K.-C.; Zhang, C. T. J. *Protein Chem.* **1993**, *12*, 709.
69. Van Waterbeemd, H. Discriminant analysis for activity prediction. In *Chemometric Methods in Molecular Design*; Van Waterbeemd, H., Manhnhold, R., Krogsgaard-Larsen, P., Timmerman, H., Eds.; Method and Principles in Medicinal Chemistry; VCH: Weinhiem, 1995; Vol. 2, pp 265–282.
70. Cabrera-Pérez, M. A.; García, A. R.; Teruel, C. F.; Álvarez, I. G.; Sanz, M. B. *Eur. J. Pharm. Biopharm.* **2003**, *56*, 197.
71. Cabrera-Pérez, M. A.; Bermejo, M. *Bioorg. Med. Chem.* **2004**, *22*, 5833.
72. Molina, E.; González-Díaz, H.; González, M. P.; Rodríguez, E.; Uriarte, E. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 515.
73. González, M. P.; González-Díaz, H.; Molina, R.; Cabrera-Pérez, M. A.; Ramos de Armas, R. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1192.
74. Ramos de Armas, R.; González-Díaz, H.; Molina, R.; González, M. P.; Uriarte, E. *Bioorg. Med. Chem.* **2004**, *12*, 4815.
75. Ramos de Armas, R.; González Díaz, H.; Molina, R.; Uriarte, E. *Biopolymers* **2005**, *77*, 247.
76. Fresht, A. *Structure and Mechanism in Protein Science: A Guide to Enzyme Catalysis and Protein Folding*; W.H. Freeman and Company: New York, 1999.
77. Esteve, V.; Blondelle, S.; Celda, B.; Perez-Paya, E. *Biopolymers* **2001**, *59*, 467.
78. Zhou, H.; Zhou, Y. *Protein Sci.* **2002**, *11*, 2714.

79. Marrero-Ponce, Y.; Medina-Marrero, R.; Castro, E. A.; Ramos de Armas, R.; González-Díaz, H.; Romero-Zaldivar, V.; Torrens, F. *Molecules* **2004**, *9*, 1124.
80. Marrero-Ponce, Y.; Medina-Marrero, R.; Castillo-Garit, J. A.; Romero-Zaldivar, V.; Torrens, F.; Castro, E. A. *Bioorg. Med. Chem.* **2005**, *13*, 3003.
81. González, M. P.; Morales, A. H.; Molina, R. *Polymer* **2004**, *45*, 2773.
82. González, M. P.; Morales, A. H.; González-Dìaz, H. *Polymer* **2004**, *45*, 2073.
83. Morales, A. H.; González, M. P.; Rieumont, J. B. *Polymer* **2004**, *45*, 2045.
84. Wiener, H. *J. Am. Chem. Soc.* **1947**, *69*, 17.
85. Estrada, E. *Chem. Phys. Lett.* **2001**, *336*, 248.
86. González-Díaz, H.; Hernández, S. I.; Uriarte, E.; Santana, L. *Comput. Biol. Chem.* **2003**, *27*, 217.
87. González-Díaz, H.; Ramos de Armas, R.; Molina, R. *Bull. Math. Biol.* **2003**, *65*, 991.
88. González-Díaz, H.; Torres-Gómez, L. A.; Guevara, Y.; Almeida, M. S.; Molina, R.; Castañedo, N.; Santana, L.; Uriarte, E. *J. Mol. Mod.* **2005**, *11*, 116.
89. González-Díaz, H.; Agüero-Chapin, G.; Cabrera-Pérez, M. A.; Molina, R.; Santana, L.; Uriarte, E.; Delogu, G.; Castañedo, N. *Bioorg. Med. Chem.* **2005**, *15*, 551.
90. González-Díaz, H.; Cruz-Monteagudo, M.; Molina, R.; Tenorio, E.; Uriarte, E. *Bioorg. Med. Chem.* **2005**, *13*, 1119.
91. González-Díaz, H.; Zaís-Urra, L.; Molina, R.; Uriarte, E. *Polymer* **2005**, *46*, 2791.