Available online at www.sciencedirect.com

SCIENCE DIRECT®

Bioorganic & Medicinal Chemistry 13 (2005) 323–331

Bioorganic & Medicinal Chemistry

Predicting stability of Arc repressor mutants with protein stochastic moments

Humberto González-Díaz,[a,b,*] Eugenio Uriarte[a] and Ronal Ramos de Armas[b]

[a]*Department of Organic Chemistry, Faculty of Pharmacy, University of Santiago de Compostela 15706, Spain*
[b]*Chemical Bioactives Center, Central University of 'Las Villas' 54830, Cuba*

Received 15 September 2004; revised 8 October 2004; accepted 9 October 2004

Abstract—As more and more protein structures are determined and applied to drug manufacture, there is increasing interest in studying their stability. In this study, the stochastic moments ($^{SR}\pi_k$) of 53 Arc repressor mutants were introduced as molecular descriptors modeling protein stability. The Linear Discriminant Analysis model developed correctly classified 43 out of 53, 81.13% of proteins according to their thermal stability. More specifically, the model classified 20/28 (71.4%) proteins with near wild-type stability and 23/25 (92%) proteins with reduced stability. Moreover, validation of the model was carried out by re-substitution procedures (81.0%). In addition, the stochastic moments based model compared favorably with respect to others based on physicochemical and geometric parameters such as D-Fire potential, surface area, volume, partition coefficient, and molar refractivity, which presented less than 77% of accuracy. This result illustrates the possibilities of the stochastic moments' method for the study of bioorganic and medicinal chemistry relevant proteins.

1. Introduction

In general, the search of novel molecular descriptors in the range of small-to-medium sized molecules in order to seek quantitative-structure–activity-relationships (QSAR)[1] constitutes nowadays a widely covered field with more than 1000 molecular descriptors introduced.[2] By the contrary, the search for theoretic approaches reaching to new polymer molecular descriptors can be classified as an emerging area with a pioneer work on the radius of gyration reported by Flory in 1953.[3] More recently appeared other approaches which are potential sources, define, or apply in some extent polymer descriptors, such as Roy et al.,[4] Casanovas et al.,[5] and Leong and Mogenthaler representations;[6] the Arteca's mean over crossing number,[7,8] the 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-Dolittle hydrophobicity.[13] Some of these initial works in polymers structure codification and QSAR have made use of the concept of moment such as Gutman and Rosenfield studies on polymers' graphs,[14] mean hydrophobicity moment[13] and the I3 index for proteins.[15,16] The success of the moments method on the field of polymers has been also confirmed after Gónzalez and co-workers work this year with implications on the analysis of the mutagenic power of dental polymers,[17,18] which preceded the very related and recent work of Morales et al.[19] In general, the moments method has been largely used in many other different contexts of solid, theoretic, and bioorganic chemistry. Burdett and Lee have applied this method to the analysis of solids energy.[20–22] In the field of theoretic chemistry should be considered as seminar works on the moments method those developed by Gutman and co-workers.[23,24] In connection to theoretic chemistry topics too, Jiang et al.[25] have analyzed the implications of the moments method on the Hueckel molecular orbital theory. Another interesting application in theoretic quantum chemistry was the work by Karwowski et al.[26] on Hamiltonian matrices. One of the authors of the present work has recently co-authored a paper regarding to the limits of applicability of graph theoretic spectral moments.[27] Gónzalez and co-workers has additionally reported interesting applications of the moments method in bioorganic, medicinal chemistry,

Keywords: Stochastic descriptors; Moments; Markov chains; Protein stability; QSAR.

* Corresponding author. Tel.: +53 42 281473x131; fax: +53 42 281473x455; e-mail addresses: humbertogd@vodafone.es; humbertogd@usc.es

0968-0896/$ - see front matter © 2004 Elsevier Ltd. All rights reserved.
doi:10.1016/j.bmc.2004.10.024

and eco-toxicology of active compounds[28,29] and the molecular design of herbicides.[30] Other interesting applications of moments method in pharmaceutical sciences and bioorganic medicinal chemistry were reported by Cabrera-Pérez and co-workers, pharmacokinetics of quinolone drugs;[31,32] and Molina et al., design of antibiotics.[33] Last but not least, have appeared several applications on medicinal chemistry of the moments method developed by Estrada and co-workers, including the design of anticancer, and sedative/hypnotic compounds[34–36] and Estrada and Peña on the design of anticonvulsant drugs.[37]

However, in spite of all this broad range of applications of the moments method, which demonstrate its potentialities, applications on the field of bioorganic chemistry related to proteins science are still uncommon. Recently our group has worked on a Markov model that use entropy like descriptors to encode molecular structure with applications in toxicology,[38] nucleic acids,[39] proteins,[40–42] and bioorganic chemistry.[43,44] This method allows for an interpretation of structural moments in stochastic terms besides, which resulted in additional models predicting drugs,[45–47] nucleic acids,[48] and proteins activity too.[49]

These promising results, the above argued necessity of novel QSAR models in proteins bioorganic chemistry, and the possibility of extending the applications of the moments method aimed us to apply our stochastic model to predict the stability for a series of mutants of the protein Arc repressor. Polymers, in special proteins, must remain stable during processes such as fermentation, purification, formulation, storage, and administration to the patients as drugs.[50] Numerous researchers worldwide have worked on the development of models to predict the stability of mutants of a wild protein.[51–61] A great deal of work is currently underway to determine the contribution of individual residues to the overall fold and stability of a protein.[62] This is a very challenging problem due to the complexity of both the native and unfolded states[63] and the transition between them. Particularly, great attention have been focused on the *Arc repressors*. This protein provides an attractive system with which to address this issue because it is small (53 aa) and is amenable to genetic 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 8–14, and α-helices formed by residues 15–30 and 32–48.[64] Nevertheless, neither Zhou and Zhou's work nor other previous studies reported in the literature have attempted to predict the stability of Arc repressors.[51–61] Until our concern, our group reported by the first time a model to predict Arc mutant's stability,[42] in the work reported here we have addressed this issue using as other alternative scheme our stochastic moments approach and developed a suitable model.

2. Probabilities, entropies, and stochastic moments for electrostatic charge distribution within the protein backbone

This approach used a Markov Chain (MC)[65] model to codify information about molecular structure. A precise definition of the descriptors generated by this methodology named MARCH-INSIDE (Markovian Chemicals In Silico Design) can be found in several reports of its application in the study of several biological properties.[38–49] Briefly, we can say that MARCH-INSIDE methodology considers as states of the MC any atom, nucleotides, or aminoacids in the molecule in dependence of the kind of molecule to be described: small-to-medium sized drug, a nucleic acid or a protein, respectively.[40]

The method uses as source of molecular descriptor the $^{1}\Pi$ matrix (the one-step electron-transition stochastic

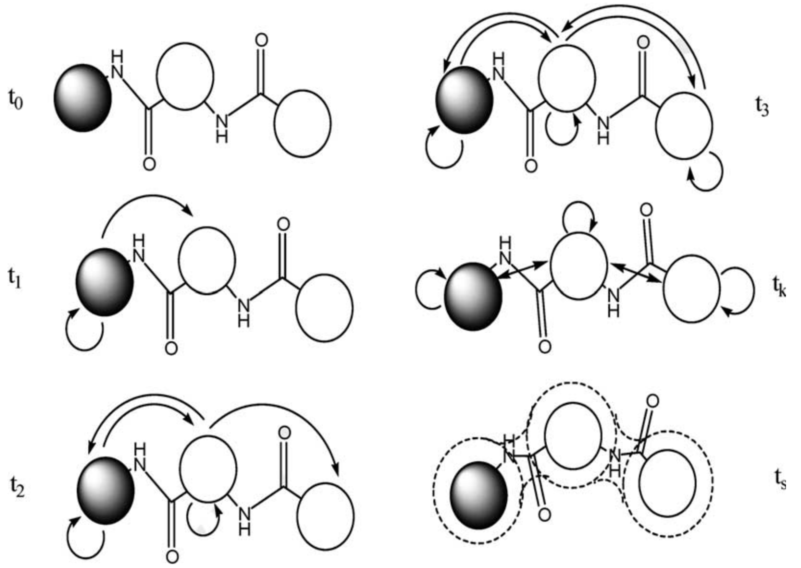


Figure 1. Representation of stochastic aa's distribution kinetic in a simple Markovian model of molecule formation. The symbol t_s, indicates stationary time: the time at which electrons reach equilibrium distribution around amino acid residues.

matrix) built up as a squared matrix $n \times n$ (n number of atoms, nucleotides or aminoacids in the molecule). Due to this work deal with proteins we used in the present definition only aminoacids, represented as aa form now on.[40–42,44,49]

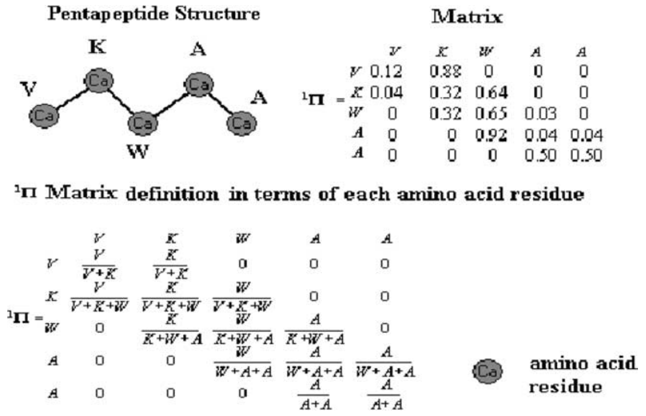
One can consider a hypothetical situation in which a set of aa residues are free in space at an arbitrary initial time (t_0). Alternatively, one can imagine a more real situation in which, after a perturbation by some external factor, electron density around these aa residues reaches a distribution different to the density distribution in the stationary state. In this case, it is of interest to develop a simple stochastic model for the distribution within the protein backbone and return of electrons to the original position with time. It can be supposed that, after this initial situation, electrons around amino acid residues begin to distribute in different ways at discrete intervals of time (t_k with $k = 0, 1, 2, \ldots$).

Thus, by using MC theory it is possible to develop a simple model of the probabilities with which the amino acid electron density changes in subsequent intervals of time until a stationary or steady state distribution arises (see Fig. 1). As depicted in Figure 1, such a model will deal with the calculation of the probabilities (${}^{k}p_{ij}$) with which the electron distributions of aa move from any aa in vicinity i at time t_0 (in black) to another aa j (in white) along discrete time periods.[38,39,42,43,45,46]

In this context, the elements (${}^{1}p_{ij}$) of ${}^{1}\Pi$ were calculated as the ratio between the electronic charge index (ECI)[66] for the jth aa and the sum of this charge over all the δ aa covalently bounded or linked through hydrogen bond to the ith aa plus 1 including itself; as exemplified Figure 2, see also Eq. 1:

$$ {}^{1}p_{ij} = \frac{\mathrm{ECI}_j}{\sum_{m=1}^{\delta+1} \mathrm{ECI}_m} \tag{1} $$

$$ {}^{A}p_0(j) = \frac{\mathrm{ECI}_j}{\sum_{m=1}^{n} \mathrm{ECI}_m} \tag{2} $$



$$ {}^{1}\Pi = \begin{array}{c|ccccc} & V & K & W & A & A \\ \hline V & 0.12 & 0.88 & 0 & 0 & 0 \\ K & 0.04 & 0.32 & 0.64 & 0 & 0 \\ W & 0 & 0.32 & 0.65 & 0.03 & 0 \\ A & 0 & 0 & 0.92 & 0.04 & 0.04 \\ A & 0 & 0 & 0 & 0.50 & 0.50 \end{array} $$

$$ {}^{1}\Pi = \begin{array}{c|ccccc} & V & K & W & A & A \\ \hline V & \frac{V}{V+K} & \frac{K}{V+K} & 0 & 0 & 0 \\ K & \frac{V}{V+K+W} & \frac{K}{V+K+W} & \frac{W}{V+K+W} & 0 & 0 \\ W & 0 & \frac{K}{K+W+A} & \frac{W}{K+W+A} & \frac{A}{K+W+A} & 0 \\ A & 0 & 0 & \frac{W}{W+A+A} & \frac{A}{W+A+A} & \frac{A}{W+A+A} \\ A & 0 & 0 & 0 & \frac{A}{A+A} & \frac{A}{A+A} \end{array} $$

Figure 2. ${}^{1}\Pi$ matrix calculation, where the symbol of the aa indicates its ECI value, for example: A is the ECI for alanine.

Also a new matrix (vector matrix), the ${}^{A}\Pi_k$ matrix, can be defined as the product of a $1 \times n$ vector (${}^{A}\Pi_0$) and the ${}^{k}\Pi$ matrix, which is the kth power of ${}^{1}\Pi$ matrix. The elements (${}^{A}p_0(j)$) of ${}^{A}\Pi_0$ were calculated in the similar way as ${}^{1}p_{ij}$ but summing up the charge of all the aa in the protein, see Eq. 2.

These matrices can then be used to generate three families of molecular descriptors:

(a) Absolute probabilities of the movement of electrons within the protein backbone (${}^{A}\pi_k(j)$) are the elements of the ${}^{A}\Pi_k$ matrix:[38,40,42,44]

$$ {}^{A}\Pi_k = {}^{A}\Pi_0 \times {}^{k}\Pi = {}^{A}\Pi_0 \times ({}^{1}\Pi)^k \tag{3} $$

Codify the attraction of the jth aa over any electron in the protein at any time t_k after traveling by different paths composed by less than k steps or aa. The ${}^{A}\pi_k(j)$ were referred in this work only as elements necessary to calculate the Electronic delocalization entropies Θ_k.

(b) Electronic delocalization entropy or Markovian negentropies (Θ_k):[38,40,42,44]

$$ \Theta_k = -k_{\mathrm{B}} \sum_{j=1}^{n} {}^{A}\pi_k(j) \log {}^{A}\pi_k(j) \tag{4} $$

where k_{B} is the Boltzman constant. The Θ_k describes the entropy involved in the electron attraction at least k steps or the same at least k aa beginning with the j aa.

(c) Stochastic moments or self return probabilities (${}^{\mathrm{SR}}\pi_k$):[45–49]

$$ {}^{\mathrm{SR}}\pi_k = \mathrm{Tr}({}^{k}\Pi) = \mathrm{Tr}[({}^{1}\Pi)^k] = \sum_{i=1}^{n} {}^{k}p_{ii} \tag{5} $$

It can be defined as the trace (sum over the ${}^{k}p_{ii}$ values) of the kth power of the ${}^{1}\Pi$ matrix. Codify the attraction of an aa for its own electrons located at kth steps away or less within the protein backbone at any time t_k after the initial time t_0.

3. Results and discussion

MC models are well-known tools for analyzing biological sequence data and they have been used to find new genes from the open reading frames.[67,68] Another use of these models is data-based searching and multiple sequence alignment of protein families and protein domains.[69] Protein-turn types[70] and sub-cellular locations have been successfully predicted.[71,72] Hubbard and Park[73] used amino acid sequence-based hidden Markov Models to predict secondary structures. In this sense, Krogh et al.[69] have also proposed a hidden Markov Model architecture. In addition, Markov's stochastic process has been used for protein folding recognition.[74] This approach can also be used for the prediction of protein signal sequences.[75,76] Another seminar works after Chou[77,78] can be found related to the application of MC theory to Proteomic and Bioinformatics. This author has also applied MC models to predict beta turns and their types,[79] and the prediction of protein cleavage sites by HIV protease.[80]

In this work we used a series of MC stochastic moments (${}^{\mathrm{SR}}\pi_k$) were used as molecular descriptors to discriminate between stable and unstable Arc repressor mutants. These aforementioned descriptors, which are invariants of a MC model for electrons delocalization in protein backbones, are used to seek the QSAR. The best equation found after LDA analysis was:

$$\begin{aligned}&\text{Stability} = 3.667 \times {}^{\mathrm{SR}}\pi_1 - 3.213 \times {}^{\mathrm{SR}}\pi_2 + 33.348\\ &N = 53 \quad \lambda = 0.63 \quad F(2,50) = 14.5\\ &p < 0.001 \quad \% = 81.1 \quad \%^{+} = 71.4 \quad \%^{-} = 92.0 \qquad (6)\end{aligned}$$

where N is the number of proteins used in the study including alanine-mutants and the wild-type Arc repressor (wtArc). The statistical parameters of the above equation are also shown and include Wilk's statistic (λ), Fischer ratio (F), and significance level (p).[81,82] 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 81.13%. More specifically, the model classified 20/28 proteins with near wild-type (nwt) stability (71.4%$^{+}$) and 23/25 (92%$^{-}$) proteins with decreased stability (ds). Table 1 depicted the respective classification matrices for training as well as cross-validation.

A cross-validation procedure was subsequently performed in order to assessing model predictability. This cross-validation was carried out by means of a resubstitution technique composed by some main stages. First, we select out of training series at random 25% of the compounds and constitute a predicting series (cv1). Afterwards, compounds in predicting series are interactively interchanged with those in training ones (cv2, cv3, cv4). Finally, one reports the accuracy, and classification matrices for each stage, and mean or averaged results.[83] The present model has shown a quite good average predictability of 81%. In particular, the model showed a very high average accuracy of 92.1% predicting the stability class of decreased stability mutants. Near-wild-type mutants are predicted with slightly low average accuracy (71.2), which is however a significant result. The importance of this result relates also to the simplicity of the present topologic methodology, which do not need 3D exhaustive structural information. The present result coincides with other reported by others with respect to the possibilities of different structural-spectral-moment-based approaches in polymer sciences. Tables 2 and 3 depicted detailed results for each protein, in connection to observed classification versus predicted training, and cross-validation, in both classes.

It is remarkable that, the predicted stability of almost proteins do not diverge from one partition to other. That is to say, the classification of a protein do not depend on the specific proteins used to conform the training or the predicting series, which points to a higher model stability under variation of data. One exception is 28EA ST11, this protein is on average misclassify as reduced stability protein but was correctly classified by the model trained with partition 1 (p1). However, it seems that there is not structural apparent reason to classify this protein or any other as statistical outlier in the present study.

Finally, may be of the major interest the physical interpretation of Eq. 6. In order to facilitate this we express this model as a function not exactly of the absolute value of for a given mutant but as a function of the difference of these values with respect to the wtArc (${}^{\mathrm{SR}}\pi_k$). Consequently, the positive influence of ${}^{\mathrm{SR}}\pi_1$ express that the higher the probability of stabilizing short-range interactions the higher will be (in a proportion of 3.667 times) the stabilization of the mutant. Conversely, long-range interaction involving not only bound amino acid but those at distance 2 during the mutation process (${}^{\mathrm{SR}}\pi_2$) may cause destabilization. It could be determined by destabilization of short-range interactions due to the stabilization of long-range ones. Thence, there is interrelated stabilization/destabilization of short-range/long-range interactions.

In closing, we carried out a comparison of the present model with other models predicting the stability of Arc repressor mutants published by the first time very

Table 1. Classification matrices and accuracy for training and re-substitution cross-validation

Train				cv-Average			
Class	%	nwt	ds	Class	%	nwt	ds
nwt	71.4	**20**	8	nwt	71.2	**15**	6
ds	92	2	**23**	ds	92.1	2	**17**
Total	81.13			Total	81.0		
cv1				cv2			
nwt	80.0	**16**	4	nwt	66.7	**14**	7
ds	89.5	2	**17**	ds	100.0	0	**18**
Total	84.6			Total	82.1		
cv3				cv4			
nwt	71.4	15	6	nwt	66.7	**14**	7
ds	89.5	2	17	ds	89.5	2	**17**
Total	80.0			Total	77.5		

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

Table 2. Classification results for all near-wild-type (nwt) mutants

Mutant[a]	O[b]	pt[c]	cv1[d]	p1[c]	cv2[d]	p2[c]	cv3[d]	p3[c]	cv4[d]	p4[c]	pm[d]
01MA ST6	nwt	89.8		87.3	nwt	93.8	nwt	88.1	nwt	88.1	89.3
02KA ST6	nwt	88.9	nwt	85.9		93.1	nwt	87.2	nwt	87.2	88.4
03GA ST6	nwt	87.5	nwt	81.7	nwt	91.9		87.1	nwt	87.1	87.0
04MA ST6	nwt	93.9	nwt	94.0	nwt	96.8	nwt	91.2		91.2	93.3
05SA ST6	nwt	92.0		95.7	nwt	95.7	nwt	82.6	nwt	82.6	89.2
06KA ST6	nwt	92.9	nwt	95.3		96.3	nwt	86.8	nwt	86.8	91.3
07MA ST6	nwt	94.9	nwt	96.0	nwt	97.5		91.2	nwt	91.2	94.0
08PA ST6	nwt	90.9	nwt	90.6	nwt	94.7	nwt	87.6		87.6	90.1
09QA ST6	nwt	91.5		81.0	nwt	94.7	nwt	93.6	nwt	93.6	90.7
11NA ST6	nwt	95.0	nwt	90.1		97.3	nwt	95.7	nwt	95.7	94.7
13RA ST6	nwt	88.2	nwt	83.5	nwt	92.5		87.4	nwt	87.4	87.7
16RA ST6	nwt	95.1	nwt	98.7	nwt	97.8	nwt	83.1		83.1	90.6
17EA ST6	nwt	86.8		78.7	nwt	91.2	nwt	87.5	nwt	87.5	86.2
18VA ST6	nwt	89.9	nwt	87.7		93.9	nwt	87.9	nwt	87.9	89.4
20DA ST6	nwt	85.6	nwt	97.6	nwt	91.9		49.9	nwt	49.9	72.3
[*]23RA ST11	nwt	20.9	nwt	27.7	nwt	19.9	nwt	12.9		12.9	18.3
[*]24KA ST11	nwt	14.3		6.0	nwt	11.4	nwt	20.2	nwt	20.2	14.5
25VA ST6	nwt	90.4	nwt	88.6		94.2	nwt	88.1	nwt	88.1	89.8
27EA ST6	nwt	93.1	nwt	92.4	nwt	96.2		90.8	nwt	90.8	92.6
[*]28EA ST11	nwt	30.8	nwt	68.9	nwt	33.8	nwt	8.9		8.9	30.1
[*]34NA ST11	nwt	13.9		6.9	nwt	11.1	nwt	17.8	nwt	17.8	13.4
35SA ST6	nwt	95.7	nwt	98.3		98.1	nwt	88.4	nwt	88.4	93.3
[*]39QA ST11	nwt	22.4	nwt	23.4	nwt	21.1		17.0	nwt	17.0	19.6
43EA ST6	nwt	79.0	nwt	73.8	nwt	84.6	nwt	76.6		76.6	77.9
[*]46KA ST11	nwt	16.6		7.3	nwt	13.7	nwt	22.8	nwt	22.8	16.6
[*]52GA ST11	nwt	12.6	nwt	5.0		9.7	nwt	18.5	nwt	18.5	12.9
[*]ARC ST11	nwt	17.3	nwt	8.8	nwt	14.6		21.7	nwt	21.7	16.7
ARC ST6	nwt	89.6		87.1	nwt	93.7	nwt	87.7		87.7	89.1

[*] Misclassified protein in training series.

[a] Arc repressor mutant code: position of mutation, specific mutation, and terminal peptide, for example, 03GA ST6 refers to a mutant of Arc, which in position 03 changes glycine (G) to alanine (A) and was coupled with the terminal peptide ST6.

[b] Observed mutant stability: nwt indicates near wild-type stability and ds points to decreased stability with respect to wild-type Arc repressor.

[c] Resulting probability (%) with which a mutant is classified in training (pt), four different re-substitution cross-validation experiments pcv1, pcv2, pcv3, pcv4, and the mean value for cross-validation (pm).

[d] Classification in training and cross-validation series (nwt) near wild type protein, (ds) decreased stability protein, ( ) leave-out protein.

recently by our group.[42] As depicted in Table 4 in general one-descriptors models based in classic physicochemical and geometric parameters such as D-Fire potential (DF),[84] surface area (SA), volume (V), partition coefficient ($\log P$), and molar refractivity (M_R)[2,42,85] presented weak linear relationship with Arc repressors stability than the 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 presented not significant relationship with Arc repressors stability ($p > 0.05$). No significant models of two variables combining classic physicochemical and geometric parameters could be found.

We have in addition explored, using forward stepwise strategy, alternative models with one up to three variables. The model with only one variable ($^{SR}\pi_1$) is statistically significant too but has overall accuracy lower than 80% because of a fall down until 67.8% of the accuracy for nwt proteins. On the other hand, the model with three variables no offer any advantage in accuracy terms being the third variable introduced $^{SR}\pi_3$ not significant ($p = 0.29$) in statistical terms. Last but not the least, the Receiver Operating Characteristic (ROC) curve has been built up for training and predicting series. Notably, the curve presented a pronounced curvature (convexity) with respect to the $y = x$ line for both training and predicting series. This result confirm that the present model it is a significant classifier having area under ROC curve above 0.8 higher than 0.5, which is the value for a random classifier (Fig. 3).[86]

These results coincided with those reporting a better recognition of proteins folding and function by MC descriptors than the classic ones.[41] The MC model based on the concept of entropy ($\Delta\Theta_0$)[42] with only one variable performs equal than the stochastic moments model, however it is straightforward to realize that we need two kind of matrices ($^{A}\Pi_0, {}^{k}\Pi$) and a logarithmic relationship defining this entropic parameter conversely the moments are simpler additive descriptors base only on the $^{k}\Pi$ matrices. Finally, a combined model with MC and classic parameters performs slightly better than our stochastic moments model but having still higher molecular descriptors complexity.[42]

4. Conclusions

As a sort of concluding remark the present work introduces a novel method to classify Arc repressor mutants

Table 3. Classification results for all mutants with decreased stability (ds) with respect to wild-type Arc repressor

Mutant[a]	O[b]	pt[c]	cv1[d]	p1[c]	cv2[d]	p2[c]	cv3[d]	p3[c]	cv4[d]	p4[c]	pm[d]
*10FA ST6	ds	8.8	ds	8.1		5.0	ds	13.1	ds	13.1	9.8
12LA ST11	ds	81.1	ds	88.9	ds	83.6		78.4	ds	78.4	82.4
14WA ST11	ds	84.5	ds	93.2	ds	87.4	ds	78.7		78.7	84.5
15PA ST11	ds	79.2		87.0	ds	81.5	ds	77.1	ds	77.1	80.7
*19LA ST6	ds	11.5	ds	15.1		7.2	ds	13.0	ds	13.0	12.1
21LA ST11	ds	83.6	ds	91.3	ds	86.4		80.0	ds	80.0	84.4
22VA ST11	ds	75.5	ds	83.9	ds	77.3	ds	73.9		73.9	77.3
29NA ST11	ds	59.3		52.1	ds	56.7	ds	71.4	ds	71.4	62.9
30GA ST11	ds	83.3	ds	91.8		86.1	ds	78.6	ds	78.6	83.8
31RA ST11	ds	79.5	ds	89.5	ds	82.1		74.5	ds	74.5	80.2
32SA ST11	ds	79.5	ds	89.5	ds	82.1	ds	74.5		74.5	80.2
33VA ST11	ds	82.0		90.5	ds	84.7	ds	78.0	ds	78.0	82.8
36EA ST11	ds	73.0	ds	56.7		72.5	ds	86.4	ds	86.4	75.5
37IA ST11	ds	81.6	ds	90.0	ds	84.2		77.8	ds	77.8	82.5
38YA ST11	ds	73.1	ds	83.7	ds	74.8	ds	69.8		69.8	74.5
40RA ST11	ds	68.3		29.9	ds	65.1	ds	90.8	ds	90.8	69.2
41VA ST11	ds	82.2	ds	90.7		84.9	ds	78.1	ds	78.1	82.9
42MA ST11	ds	75.2	ds	83.4	ds	77.0		73.9	ds	73.9	77.1
44SA ST11	ds	74.0	ds	73.0	ds	74.7	ds	80.0		80.0	76.9
45FA ST11	ds	78.9		87.2	ds	81.2	ds	76.5	ds	76.5	80.3
47KA ST11	ds	80.5	ds	87.7		82.8	ds	78.7	ds	78.7	82.0
48EA ST11	ds	66.2	ds	68.7	ds	65.9		71.2	ds	71.2	69.3
49GA ST11	ds	83.3	ds	91.9	ds	86.1	ds	78.6		78.6	83.8
50RA ST11	ds	83.7		81.7	ds	85.5	ds	88.5	ds	88.5	86.0
51IA ST11	ds	83.1	ds	91.7		85.9	ds	78.5	ds	78.5	83.6

[*] Misclassified protein in training series.

[a] Arc repressor mutant code: position of mutation, specific mutation, and terminal peptide, for example, 03GA ST6 refers to a mutant of Arc, which in position 03 changes glycine (G) to alanine (A) and was coupled with the terminal peptide ST6.

[b] Observed mutant stability: nwt indicates near wild-type stability and ds points to decreased stability with respect to wild-type Arc repressor.

[c] Resulting probability (%) with which a mutant is classified in training (pt), four different re-substitution cross-validation experiments pcv1, pcv2, pcv3, pcv4, and the mean value for cross-validation (pm).

[d] Classification in training and cross-validation series (nwt) near wild type protein, (ds) decreased stability protein, ( ) leave-out protein.

Table 4. Results of the comparative study of the present approach with respect to the other five stability scoring functions

	Physicochemical and geometric parameters					Combined	MC-models	
Stat[a]	DF[b]	SA[c]	V[d]	Log$P$[e]	M_R[f]	$\Delta\Theta_0$,DF	$\Delta\Theta_0$	$^{SR}\pi_1$, $^{SR}\pi_2$
%T	76.9	70.7	62.3	59.0	60.0	82.7	81.1	81.1
%nwt	92.9	63.6	53.6	80.8	77.3	70.0	71.4	71.4
%ds	58.3	78.9	72.0	15.4	38.9	96.0	92.0	92
%T_{cv}	71.8	61.5	56.4	48.7	61.5	81.2	79.5	81.2
N	53	53	53	53	53	53	53	53
λ	0.79	0.85	0.92	0.99	0.97	0.46	0.56	0.63
F	13.9	8.8	4.2	0.5	1.8	11.60	39.05	14.5
p	0.0	0.0	0.0	0.5	0.2	0.00	0.00	0.00

Forward stepwise analysis

Variables	N	%T	%nwt	%ds	λ	F	p-Last[g]
$^{SR}\pi_1$	53	79.2	67.8	92.0	0.68	23.9	0.00
$^{SR}\pi_1$, $^{SR}\pi_2$	53	81.1	71.4	92.0	0.63	14.5	0.00
$^{SR}\pi_1$, $^{SR}\pi_2$, $^{SR}\pi_3$	53	81.1	71.4	92.0	0.61	10.1	0.29

[a] Statistic parameters verifying model quality: %T, %nwt, %ds, %T_{cv} are the total, near wild-type group, decreased stability group, and average cross-validation 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] p-Last is the p-level for the last variable entered in the model.

with respect to its stability. The paper makes emphasis on expanding the possibilities of the method of moments on proteins and bioorganic medicinal chemistry coinciding with very recent results after González and Terán.[87–89]

The model here reported based on stochastic moments performs better than other models based on physicochemical and geometric descriptors acting as simple descriptors or in groups of two variables coinciding with

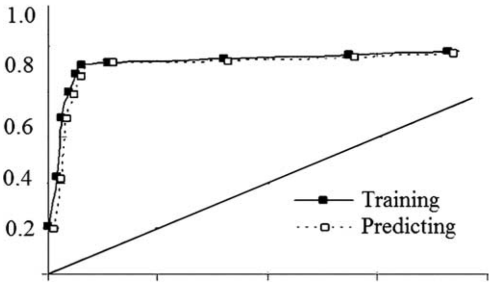


Figure 3. Operating receive characteristic curve (ROC-curve) for training and predicting series of antibacterial and nonactive compounds.

other finding previously reported by our group. At the same time, this work confirms the potentialities of stochastic molecular indices.

5. Experimental data and analysis

The Arc repressor mutant data was taken from the literature.[64] Alanine mutations were constructed for the 51 nonalanine positions and each mutant was then purified and subjected to thermal and urea denaturation. The melting temperature (T_m) was determined in order to check the stability of the protein.

Two groups were built in order to perform LDA analysis: proteins with near wild-type stability ($T_m > 53\,°C$) and proteins that decreased stability ($T_m < 53\,°C$). The ECI values were also taken from the literature.[64]

The protein backbone of the homodimer was built using the '*draw mode*' of the program MARCH-INSIDE 1.0.[90] In this respect we only considered covalent interactions (peptidic bond) and hydrogen bonding interactions (within a chain as well as between chains). As a first approximation, we considered both interactions as being equivalent, taking into account the 'connectivity of the protein'. The mutants were then constructed by changing an aa for alanine and considering that this change only affects the possibility in this region for the protein to form polar interactions (the hydrogen interaction was suppressed if the former aa had such an interaction). Finally, the first five molecular descriptors were calculated (${}^{SR}\pi_1$ to ${}^{SR}\pi_5$) in order to perform LDA analysis. The starting point of the Markov chain ${}^{SR}\pi_0$, which is equal to the number of amino acids in the protein n, was ignored to avoid overestimation of the effect of polypeptide tails considered as not affecting protein stability.[42,64]

Linear discriminant analysis (LDA)[81–83] was done in order to classify into two groups. The melting temperature (T_m) was taken into consideration in order to check the stability of the protein and define the two groups to perform LDA analysis: proteins with near wild-type stability ($T_m > 53\,°C$) and proteins that decreased stability ($T_m < 53\,°C$). For the LDA analysis, we employed the linear discriminant analysis module of the STATISTICA software. Previously to LDA analysis data was mean centered, standardized, and the Randič's orthogonalization was applied avoiding variables collinearity.[91–93] The assessment of the statistical quality of the models was done by some well-known parameters such as Wilk_s lambda (k), Fischer ratio (F), squared Mahlanobis distance (D^2) and the percentage of good classification for the training set as well as for cross validation procedure.

Acknowledgements

González-Díaz, H. thanks the University of 'Santiago de Compostela' in Spain, for kind hospitality and the Spanish Ministry of Science and Technology (SAF2003-02222), and the Xunta de Galicia (PR405A2001/65-0) for partial financial support.

References and notes

1. Kubinyi, H.; Taylor, J.; Ramdsen, C. Quantitative Drug Design. In *Comprehensive Medicinal Chemistry*; Hansch, C., Ed.; Pergamon: Oxford, UK, 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*, 721–728.
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*, 74–81.
14. Gutman, I.; Rosenfield, V. R. *Theor. Chim. Acta* **1996**, *93*, 191.
15. Estrada, E. *Bioinformatics* **2002**, *18*, 1.
16. Estrada, E. *Chem. Phys. Lett.* **2000**, *319*, 713.
17. González, M. P.; Morales, A. H.; Molina, R. *Polymer* **2004**, *45*, 2773.
18. González, M. P.; Morales, A. H.; González-Díaz, H. *Polymer* **2004**, *45*, 2073.
19. Morales, A. H.; González, M. P.; Rieumont, J. B. *Polymer* **2004**, *45*, 2045.
20. Burdett, J. K.; Lee, S. *J. Am. Chem. Soc.* **1985**, *107*, 3063.
21. Burdett, J. K.; Lee, S. *J. Am. Chem. Soc.* **1985**, *107*, 3050.
22. Lee, S. *Acc. Chem. Res.* **1991**, *24*, 249.
23. Gutman, I. *Theor. Chim. Acta* **1992**, *83*, 313–318.
24. Markovic, S.; Gutman, I. *J. Mol. Struct. Theochem* **1991**, *81*, 81.
25. Jiang, Y.; Tang, A.; Hoffmann, R. *Theor. Chim. Acta* **1984**, *66*, 183–192.
26. Karwowski, J.; Bielinska-Waz, D.; Jurkowski, J. *Int. J. Quantum Chem.* **1996**, *60*, 185.
27. Estrada, E.; González-Díaz, H. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 75.

28. González, M. P.; González-Díaz, H.; Cabrera-Pérez, M. A.; Molina, R. R. *Bioorg. Med. Chem.* **2004**, *12*, 735.
29. González, M. P.; Morales, A. H. *J. Comput.-Aided Mol. Des.* **2003**, *10*, 665.
30. González, M. P.; González-Díaz, H.; Molina, R.; Cabrera-Pérez, M. A.; Ramos de A, R. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1192.
31. Cabrera-Pérez, M. A.; González-Díaz, H.; Fernandez, T. C.; Pla-Delfina, J. M.; Bermejo, S. M. *Eur. J. Pharm. Biopharm.* **2002**, *53*, 317.
32. Cabrera-Pérez, M. A.; García, A. R.; Teruel, C. F.; Álvarez, I. G.; Sanz, M. B. *Eur. J. Pharm. Biopharm.* **2003**, *56*, 197.
33. Molina, E.; González-Díaz, H.; González, M. P.; Rodríguez, E.; Uriarte, E. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 515.
34. Estrada, E.; Peña, A.; García-Domenech, R. *J. Comput. -Aided Mol. Des.* **1998**, *12*, 583.
35. Estrada, E.; Uriarte, E. *Curr. Med. Chem.* **2001**, *8*, 1573.
36. Estrada, E.; Uriarte, E.; Montero, A.; Teijeira, M.; Santana, L.; De Clercq, E. *J. Med. Chem.* **2000**, *43*, 1975.
37. Estrada, E.; Peña, A. *Bioorg. Med. Chem.* **2000**, *8*, 2755.
38. 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. Tox.* **2003**, *16*, 1318.
39. González-Díaz, H.; Ramos de A, R.; Molina, R. R. *Bioinformatics* **2003**, *19*, 2079.
40. González-Díaz, H.; Molina, R. R.; Uriarte, E. *Polymer* **2004**, *45*, 3845.
41. González-Díaz, H.; Molina, R. R.; Uriarte, E. *Bioorg. Med. Chem. Lett.* **2004**, *14*, 4691–4695.
42. Ramos de A, R.; González-Díaz, H.; Molina, R. R.; Uriarte, E. *Proteins: Struct., Funct., Bioinf.* **2004**, *56*, 715.
43. González-Díaz, H.; Bastida, I.; Castañedo, N.; Nasco, O.; Olazabal, E.; Morales, A.; Serrano, H. S.; Ramos de A, R. *Bull. Math. Biol.* **2004**, *66*, 1285.
44. Ramos de A, R.; González Díaz, H.; Molina, R.; González, M. P.; Uriarte, E. *Bioorg. Med. Chem.* **2004**, *12*, 4815.
45. González-Díaz, H.; Olazábal, E.; Castañedo, N.; Hernádez, S. I.; Morales, A.; Serrano, H. S.; González, J.; Ramos de A, R. *J. Mol. Model.* **2002**, *8*, 237.
46. González-Díaz, H.; Gia, O.; Uriarte, E.; Hernádez, 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. Model.* **2003**, *9*, 395.
47. González-Díaz, H.; Hernández, S. I.; Uriarte, E.; Santana, L. *Comput. Biol. Chem.* **2003**, *27*, 217.
48. González-Díaz, H.; Ramos de A, R.; Molina, R. R. *Bull. Math. Biol.* **2003**, *65*, 991.
49. González-Díaz, H.; Ramos de A, R.; Uriarte, E. *Online J. Bioinf.* **2002**, *1*, 83.
50. EUFEPS Announcement *Eur. J. Pharm. Sci.* **2002**, *15*, 101.
51. Zhou, H.; Zhou, Y. *Protein: Struct., Funct., Genet.* **2002**, *49*, 483.
52. Green, S. M.; Meeker, A. K.; Shortle, D. *Biochemistry* **1992**, *31*, 5717.
53. O'Neil, K. T.; DeGrado, W. F. *Science* **1990**, *250*, 646.
54. Blaber, M.; Zang, X.; Matthews, B. W. *Science* **1993**, *260*, 1637.
55. Kim, D. E.; Fisher, C.; Baker, D. *J. Mol. Biol.* **2000**, *298*, 971.
56. Hamill, S. J.; Steward, A.; Clarke, J. *J. Mol. Biol.* **2000**, *297*, 165.
57. Fulton, K. F.; Main, E. R. G.; Daggett, V.; Jackson, S. E. *J. Mol. Biol.* **1999**, *291*, 445.
58. Kragelund, B. B.; Osmark, P.; Neergaard, T. B.; Schidt, J.; Kristiansen, K.; Knudsen, J.; Poulsen, F. M. *Nature Struct. Biol.* **1999**, *6*, 594.
59. Ternström, T.; Mayor, U.; Akke, M.; Oliveberg, M. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 14854.
60. Lorch, M.; Mason, J. M.; Clarke, A. R.; Parker, M. J. *Biochemistry* **1999**, *38*, 1377.
61. Julenius, K.; Thulin, E.; Linse, S.; Finn, B. E. *Biochemistry* **1998**, *37*, 8915.
62. Alber, T. A. *Rev. Biochem.* **1989**, *58*, 765.
63. Dill, K. A.; Shortle, D. A. *Rev. Biochem.* **1991**, *60*, 795.
64. Milla, M. E.; Brown, M. B.; Sauer, R. T. *Struct. Biol.* **1994**, *1*, 518.
65. Freund, J. A.; Poschel, T. Stochastic Processes in Physics, Chemistry, and Biology. In *Lecture Notes in Physics*; Springer: Berlin, Germany, 2000.
66. Collantes, E. R.; Dunn, W. J. *J. Med. Chem.* **1995**, *38*, 2705.
67. Vorodovsky, M.; Koonin, E. V.; Rudd, K. E. *Trends Biochem. Sci.* **1994**, *19*, 309.
68. Vorodovsky, M.; Macininch, J. D.; Koonin, E. V.; Rudd, K. E.; Médigue, C.; Danchin, A. *Nucleic Acids Res.* **1995**, *23*, 3554.
69. Krogh, A.; Brown, M.; Mian, I. S.; Sjeander, K.; Haussler, D. *J. Mol. Biol.* **1994**, *235*, 1501.
70. Chou, K.-C. *Biopolymers* **1997**, *42*, 837.
71. Yuan, Z. *FEBS Lett.* **1999**, *451*, 23.
72. Hua, S.; Sun, Z. *Bioinformatics* **2001**, *17*, 721.
73. Hubbard, T. J.; Park, J. *Proteins: Struc., Funct., Genet.* **1995**, *23*, 398.
74. Di Francesco, V.; Munson, P. J.; Garnier, J. *Bioinformatics* **1999**, *15*, 131.
75. Chou, K.-C. *Curr. Prot. Pept. Sci.* **2002**, *3*, 615.
76. Chou, K.-C. *Peptides* **2001**, *22*, 1973.
77. Chou, K.-C. *Anal. Biochem.* **2000**, *286*, 1.
78. Chou, K.-C. *J. Biol. Chem.* **1993**, *268*, 16938.
79. Chou, K.-C. *Anal. Biochem.* **1996**, *233*, 1.
80. Chou, K.-C.; Zhang, C. T. *J. Protein Chem.* **1993**, *12*, 709.
81. Kowalski, R. B.; Wold, S. Pattern Recognition in Chemistry. In *Handbook of Statistics*; Krishnaiah, P. R., Kanal, L. N., Eds.; North Holland Publishing Company: Amsterdam, 1982; pp 673–697.
82. Cronin, M. T. D.; Aynur, A. O.; Dearden, C. J.; Deffy, C. J.; Netzeva, I. T.; Patel, H.; Rowe, H. P.; Schultz, T. W.; Worth, P. A.; Voutzolidis, K.; Schüürmann, G. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 869.
83. Van Waterbeemd, H. Discriminant Analysis for Activity Prediction. In *Chemometric Methods in Molecular Design*; Manhnhold, R., Krogsgaard-Larsen, P., Timmerman, H., Eds.; Method and Principles in Medicinal Chemistry; Van Waterbeemd, H., Ed.; VCH: Weinhiem, 1995; Vol. 2, pp 265–282.
84. Zhou, H.; Zhou, Y. *Protein Sci.* **2002**, *11*, 2714, This service is available at: http://theory.med.buffalo.edu/.
85. Fresht, A. *Structure and Mechanism in Protein Science: A Guide to Enzyme Catalysis and Protein Folding*; W.H. Freeman and Company: New York, 1999.
86. Swets, J. A. *Science* **1988**, *240*, 1285.
87. González, M. P.; Teran, M. C. M. *Bioorg. Med. Chem. Lett.* **2004**, *14*, 3077.
88. González, M. P.; Teran, M. C. M. *Bull. Math. Biol.* **2004**, *66*, 907.
89. González, M. P.; Teran, M. C. M. *Bioorg. Med. Chem.* **2004**, *12*, 2985.
90. González-Díaz, H.; Hernández, I. MARCH-INSIDE VERSION 1.0, 2002 (Markovian Chemicals 'In Silico' Design). This is

a preliminary experimental version future professional version shall be available to the public. For any information about it sends and e-mail to the corresponding author humbertogd@vodafone.es or humbertogd@usc.es.

91. Randič, M. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 311.
92. Randič, M. *New J. Chem.* **1991**, *15*, 517.
93. Randič, M. *J. Mol. Struct. (THEOCHEM)* **1991**, *233*, 45.