PROTEINS: Structure, Function, and Bioinformatics 56:715–723 (2004)

Markovian Backbone Negentropies: Molecular Descriptors for Protein Research. I. Predicting Protein Stability in Arc Repressor Mutants

Ronal Ramos de Armas,[1,2] Humberto González Díaz,[1,3]* Reinaldo Molina,[1,4] and Eugenio Uriarte[3]
[1]*Chemical Bioactives Center, Central University of "Las Villas," Cuba*
[2]*Department of Chemistry, Central University of "Las Villas," Cuba*
[3]*Department of Organic Chemistry, Faculty of Pharmacy, University of Santiago de Compostela, Spain*
[4]*Universität Rostock, FB Chemie, Rostock, Germany*

ABSTRACT **As more and more protein structures are determined and applied to drug manufacture, there is increasing interest in studying their stability. In this sense, developing novel computational methods to predict and study protein stability in relation to their amino acid sequences has become a significant goal in applied Proteomics. In the study described here, Markovian Backbone Negentropies (MBN) have been introduced in order to model the effect on protein stability of a complete set of alanine substitutions in the Arc repressor. A total of 53 proteins were studied by means of Linear Discriminant Analysis using MBN as molecular descriptors. MBN are molecular descriptors based on a Markov chain model of electron delocalization throughout the protein backbone. The model correctly classified 43 out of 53 (81.13%) 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, the model presented a good Mathew's regression coefficient of 0.643. Validation of the model was carried out by several Jackknife procedures. The method compares favorably with surface-dependent and thermodynamic parameter stability scoring functions. For instance, the D-FIRE potential classification function shows a level of good classification of 76.9%. On the other hand, surface, volume, logP, and molar refractivity show accuracies of 70.7, 62.3, 59.0, and 60.0%, respectively. Proteins 2004;56:715–723.**

INTRODUCTION

There are a huge number of pharmaceutical proteins under development. Indeed, a recent survey indicated that a large number of therapeutic monoclonal antibodies and vaccines are on their way to the market. Pharmaceutical proteins, as any proteins, are high molecular weight molecules with secondary, tertiary, and sometimes quaternary structures that are mainly stabilized by rather weak forces. These proteins must be stable during processes such as fermentation, purification, formulation, storage, and administration to the patient.[1] Moreover, knowledge of factors that determine the stability of a particular protein enables us to find out important features concerning their structure and function. In this context, the computational study of structure/stability relationships has become an important area in protein science.

Numerous researchers worldwide have worked on the development of models to predict the stability of mutants of a wild protein. For instance, Shortle et al. have studied 118 mutants of *Staphylococcal* nuclease. Similarly, other researchers have modeled the stability of 145 mutants of T4 Lysozyme, 96 mutants of Barnase, and 71 mutants of Chymotrypsin in what seem to be the models with the largest mutated proteins. Another important study involved modeling the stability of 66 mutants of GeneV, 65 mutants of Human lysozyme, and 58 mutants of protein L. Other noteworthy studies concerned 40 mutants of Trypsin inhibitor, 38 mutants of TNFn3, and 31 mutants of FKBP12. Models have also been reported for proteins with more than 10 mutants but fewer than 30, such as ACBP, Ribonuclease T1, Ribonuclease H, α Lactalbumin, Hen Lysozyme, Subtilisin inhibitor, U1A, ISO-1 cytochrome C, and Trp synthase. Other, less-mutated proteins that have been studied include CD2, Calbindin, Apomyoglobin, Adrenodoxin, Cold shock, ribonuclease A and λ-CRO. As summarized in Zhou and Zhou's excellent work, a total of 35 proteins with their respective 1023 mutants have been studied and these include all of the examples outlined above. In their review, Zhou and Zhou not only provide an excellent overview of this field but also use the data from the 1023 mutant stability tests to develop what seems to be one of the largest unified models to date.[2]

A great deal of work is currently underway to determine the contribution of individual residues to the overall fold and stability of a protein.[3] This is a very challenging problem due to the complexity of both the native and unfolded states[4] and the transition between them. Bob Sauer has done some of the seminal work in this area on

*Correspondence to: Humberto González Díaz, UCLV, Drug Design, Chermicals Bioactives Center, Central University of Las Villas, Santa Clara, 54830, Cuba. E-mail: humbertogd@cbq.uclv.edu.cu

Received 3 September 2003; Accepted 9 February 2004

Published online 20 May 2004 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.20159

the *Arc repressors.*[5] 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.[5] Nevertheless, neither Zhou and Zhou's work nor other studies reported in the literature have attempted to predict the stability of Arc repressors. In the work reported here we have addressed this issue and developed a suitable model.

The use of molecular descriptors to derive structure/property relationships is an approach of major interest.[6] Molecular descriptors are numerical indices that codify either molecular or macromolecular structure.[7] One of the most notable common molecular descriptors is represented by the topological indices.[8–10] However, physical interpretation is one of the aspects that has provided a basis for criticism of the use of topological indices.[11,12] Nevertheless, necessity drives the constant development of novel molecular indices.[13,14] In this context, some specific and very successful indices (for small molecules) that use the concept of Shannon's entropy[15] from the point of view of the information theory have proven to be very effective in drug design.[16] For example, in the 1980's L. B. Kier used the concept of entropy to codify molecular structure through the so-called Molecular Negentropy in QSAR/QSPR (Quantitative-Structure-Activity or Property-Relationship) studies.[17] Due to its general scope and special properties, the concept of entropy has been used in many other fields of science.[18] Recently, Graham analyzed the use of the Entropy concept and Information theory applied to organic molecules through Integer statistics.[19]

The MARCH-INSIDE (**Mar**kovian **Ch**emicals **In Si**lico **De**sign) methodology has been developed by our research group to generate molecular descriptors based on the Markov Chain Theory. This approach has been successfully employed in QSPR and QSAR studies, including studies related to Proteomics and Nucleic Acid-Drug interactions. The approach describes changes in the electron distribution and vibrational decay with time throughout the molecular backbone. The method allowed us to introduce physically meaningful stochastic graph invariants for the study molecular properties. The method has also demonstrated flexibility in relation to many different problems. One of the applications involved the prediction of the fluckicidal activity of novel drugs (flukes are tiny intestinal parasites).[20] More recently, the MARCH-INSIDE approach has been applied to the fast-track experimental discovery of novel anticancer compounds.[21] Additionally, promising results have been found in the modeling of the interaction between drugs and HIV-packaging-region RNA in the field of bioinformatics.[22] An alternative formulation of our approach in terms of negentropies gives more physical sense to our models for drug-RNA interactions.[23] The prediction of the biological activities of peptides and NMR shifts in proteins are problems that can also be addressed using this approach.[24,25] Codification of chirality and other 3D structural features constitutes another advantage of this method.[26] The latter opportunity has allowed the estimation of the level of agranulocytocis that is chemically induced by drugs.[27]

Markov models are well-known tools for analyzing biological sequence data and they have been used to find new genes from the open reading frames.[28,29] Another use of these models is data-based searching and multiple sequence alignment of protein families and protein domains.[30] Protein α-turn types[31] and sub-cellular locations have been successfully predicted.[32,33] Hubbard and Park[34] used amino acid sequence-based hidden Markov Models to predict secondary structures. In this sense, Krogh et al.[35] have also proposed a hidden Markov Model architecture. In addition, Markov's stochastic process has been used for protein folding recognition.[36] This approach can also be used for the prediction of protein signal sequences.[37, 38] Another seminal works can be found related to the application of Markov Chain Theory to Proteomic and Bioinformatics. Chou applied Markov Models to predict beta turns and their types,[39] and the prediction of protein cleavage sites by HIV protease.[40–42]

In this paper, we attempt to further extend this methodology to encompass protein stability studies—specifically how alanine substitution mutation on Arc repressor wild-type protein affects protein stability–by means of Linear Discriminant Analysis (LDA).

METHODS

Modeling Vibrational Delocalization Throughout the Protein Backbone

In the work described here, the MARCH-INSIDE[20–27] methodology was generalized to allow the study of Protein stability. This approach used MCH[36,42] to codify information about molecular structure. The procedure considered as states of the MCH the amino represents the number of amino acids in the protein molecule.

The elements of $^1\Pi$ (^{1}pij) were defined to codify information about the charge acids in the protein polypeptidic backbone. Accordingly, the $^1\Pi$ matrix (with elements pij) was used as a source of molecular descriptors. This matrix was called the 1-step Charge-transition stochastic matrix. $^1\Pi$ was built as a squared table of order n, where n distribution around the amino acid environment. With this aim in mind, our preliminary approach was generalized to derive molecular descriptors that helped us to predict the influence of specific alanine-amino acid substitution on protein stability. The elements (^{1}pij) of the 1-step electron-transition stochastic matrix are the transition probabilities [see Equation (1)] if the aa i and j are adjacent and are 0 otherwise:

$$ {}^1p_{ij} = \frac{\mathrm{ECI}_j}{\sum_{k=1}^{\delta+1} \mathrm{ECI}_k} \qquad (1) $$

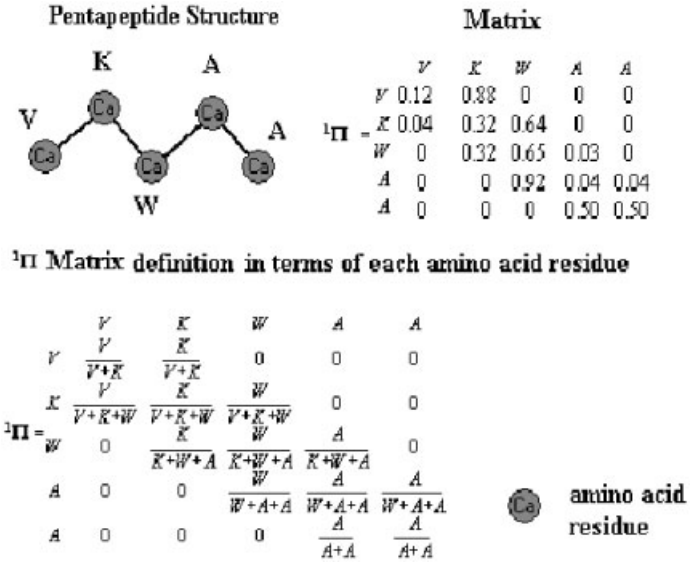

Fig. 1. $^{1}\Pi$ matrix calculation.

ECI is the Electronic Charge Index[44] descriptor for the side-chain local polarity of each amino acid. The sum is carried out over all amino acids that interact by covalent or hydrogen-bonding interactions with aa_i (including aa_i). Figure 1 exemplifies the calculation of $^{1}\Pi$ for the pentapeptide VKWAA.

This picture shows that $^{1}p_{ii}$ varies in the order: $^{1}p_{ii}$ (W) = 0.65 > (A1) = 0.50 > (K) = 0.32 > (V) = 0.12 > (A2) = 0.04. It may be concluded that $^{1}p_{ii}$ does not vary in the same order as the ECI values of each amino acid residue (W = 1.08 > K=0.53 > V = 0.07 > A2 = 0.05 = A1). The $^{1}p_{ii}$ A1 (a terminal Ala residue) value is greater than A2 (non-terminal Ala residue) in spite of the use of the same ECI value, thus resulting in efficient differentiation between amino residues by considering the topological character of $^{1}\Pi$. Figure 1 shows that $^{1}p_{ii}$ (A2) or $^{1}p_{ii}$ (A1) have identical numerators (A) but different denominators due to the different bonding of the two amino acid residues. For example, A1 is bonded to W and A2, while the A2 atom is bonded only to A1. A closer look at the $^{1}p_{ii}$ values reveals the basis of this methodology. For instance, the electrons from the non-terminal alanine will move to the adjacent tryptophan with a higher probability because of tryptophan's higher ECI value compared to the other adjacent alanine. The data suggest that the molecular indices $^{SR}\pi_k$ calculated by MARCH-INSIDE have codified electronic and topological information about the molecular structure.[20–27]

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 amino acid 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 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 = 0, 1, 2. . ., k). Thus, by using MCH [20–27,43,45,46] 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. 2). As depicted in Figure 2, such a model will deal with the calculation of the probabilities ($^{k}p_{ij}$) with which the electron distributions of amino acids move from any amino acid in vicinity i at time t_0 (in black) to another amino acid j (in white) along discrete time periods t_k (k = 1, 2, 3. . .) throughout the

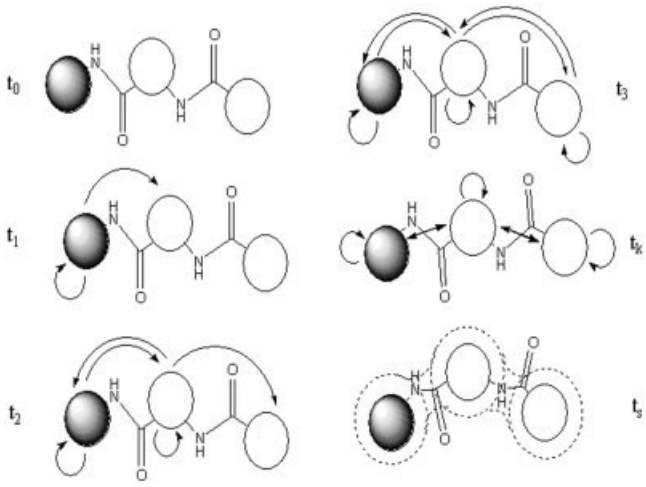

Fig. 2. Representaion of stochastic amino acids' distribution kitetic in a simple Markovian model of molecule formation. The symbol t_s indicates stationary time: the time at which electrons reach equilibrium distribution around amino acid residues.

chemical bonding system. Such a model is stochastic per se (probabilistic distribution of aa's with time) but in fact considers molecular connectivity (electron distribution in space throughout the chemical bonding system).

The Markovian Vibrational Entropies

Markovian Molecular Negentropies generalized for protein backbone molecular descriptors (Θ_k) were defined as the Entropies of the charge distribution over the whole protein molecule with time (k):

$$\Theta_k = -\sum {}^{A}p_k(j)\log({}^{A}p_k(j)) \quad (2)$$

The parameter k is neither a pure time measurement nor a topological distance coordinate with respect to j. This parameter codifies magnitudes of both time and space. The parameter k accounts for the integer intervals of time at which the intensity of the charge distribution varies with Markovian probabilities along the protein. These molecular descriptors for the protein backbone could be interpreted as the entropy involved in the charge distribution over the protein domains after time k.

The calculation of the absolute probabilities was straightforward from classical results from Markov chains theory.[20–27,43,45,46]

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

Where ${}^{A}\Pi_k$ are $1 \times n$ vectors whose elements ${}^{A}p_k(j)$ are the aforementioned absolute probabilities, ${}^{A}\Pi_0$ is a $1 \times n$ vector whose elements are the ${}^{A}p_0$ (j) probabilities for n atoms in the molecule and ${}^{k}\Pi$ are the k^{th}-natural powers of the ${}^{1}\Pi$ matrix. The ${}^{A}p_k(j)$ values were defined similarly to the ${}^{k}p_{ij}$ probabilities [Equation (1)] but consider in the sum all the amino acids in the protein molecule; see González et al. for an exhaustive explanation of this topic.[20–27] All the calculations were carried out using our experimental software MARCH-INSIDE.[47]

Experimental Data and Analysis

The Arc repressor mutant data was taken from the literature.[5] Alanine mutations were constructed for the 51 non-alanine 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[44] and are shown in Table I.

The protein backbone of the homodimer was built using the "*draw mode*" of the program. 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_i 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 10 molecular descriptors were calculated (Markovian Vibrational Entropy and Self Return Probability) in order to perform LDA analysis.

TABLE I. ECI Values of Natural Acid Side Chain Proposed to Obtain the Descriptors

aa	ECI	Δ p (ECI)	aa	ECI	Δ p (ECI)
Gly	0.02	−0.03	Asn	1.31	1.26
Ala	0.05	0	Glu	1.31	1.26
Val	0.07	0.02	Gln	1.36	1.31
Leu	0.1	0.05	Ser	0.56	0.51
Ile	0.09	0.04	Thr	0.65	0.6
Phe	0.14	0.09	Cys	0.15	0.1
Tyr	0.72	0.67	Met	0.34	0.29
Trp	1.08	1.03	Lys	0.53	0.48
Pro	0.16	0.11	Arg	1.69	1.64

For comparison purposes several parameters related to surface-area[2] and thermodynamics were calculated using HYPERCHEM[48] software. Distance–based potential (D-FIRE)[49] was also applied to the specific problem of stability prediction. The 3D structure of the Arc repressor was taken from the Protein Data Bank[50] in .ent format (pdb1arr file). This file was processed with Hyperchem software as follows: After opening the file, Mutants were built from wild-type Arc repressor dimer using the mutate command. No further structure minimization was performed on the protein.[49,51]

RESULTS

Markovian Molecular Negentropies Modeling Protein Stability

The best equation found after LDA analysis was:

$$Stability = 92.47 \times \Delta\Theta_0(Mut - Arc) - 177.85$$
$$N = 53 \quad \lambda = 0.56 \quad F(1,51) = 39.05$$
$$p < 0.000 \quad C = 0.64 \quad (4)$$

Where $\Delta\ \Theta_0(Mut - Arc)$ is the difference in entropy between an alanine mutant and the wild-type Arc suppressor at an initial time after perturbation. The statistical parameters of the above equation are also shown and include Wilk's statistic (λ), Fischer Ratio (F), Mathew's coefficient (C), and significance level (p). The discriminant function classified correctly 43 out of 53 mutant proteins according to their relative stability related to wild-type protein. These give a level of accuracy of 81.13%. More specifically, the model classified 20/28 proteins with near wild-type stability (71.4%) and 23/25 (92%) proteins with decreased stability. Table II shows the resulting probability of the classified mutant proteins as well the probabilities after a leave-one-out cross validation was carried out.

A leave-n-out procedure[52–54] was subsequently performed. In this figure the percentage of good classification is plotted versus the number of compounds extracted from the training series, as well as the cross validation error

TABLE II. Results of Classification and Leave-One-Out Cross-Validation From Equation (4)

Mutant[a]	Obs[b]	P%[c]	Pcv%[d]	Mutant[a]	Obs[b]	P%[c]	Pcv%[d]
01MA ST6	Nwt	88.06	88.43	29NA ST11	Rs	−63.03	−62.15
02KA ST6	Nwt	89.77	90.11	30GA ST11	Rs	−81.17	−80.41
03GA ST6	Nwt	78.82	79.49	31RA ST11	Rs	−67.41	−66.52
04MA ST6	Nwt	88.06	87.56	32SA ST11	Rs	−67.41	−66.52
05SA ST6	Nwt	89.95	89.55	33VA ST11	Rs	−78.57	−77.76
06KA ST6	Nwt	89.77	89.36	*34NA ST11	Nwt	−63.03	−68.18
07MA ST6	Nwt	88.06	87.56	35SA ST6	Nwt	89.95	89.55
08PA ST6	Nwt	84.78	84.15	36EA ST11	Rs	−63.03	−62.15
09QA ST6	Nwt	90.34	89.96	37IA ST11	Rs	−77.67	−76.85
*10FA ST6	Rs	84.23	89.99	38YA ST11	Rs	−62.87	−61.98
11NA ST6	Nwt	90.5	90.12	*39QA ST11	Nwt	−63.46	−68.64
12LA ST11	Rs	−77.2	−76.42	40RA ST11	Rs	−67.41	−66.52
13RA ST6	Nwt	88.73	88.26	41VA ST11	Rs	−78.57	−77.76
14WA ST11	Rs	−61.8	−60.9	42MA ST11	Rs	−69.32	−68.44
15PA ST11	Rs	−74.9	−74.03	43EA ST6	Nwt	90.5	90.12
16RA ST6	Nwt	88.73	88.26	44SA ST11	Rs	−64.83	−63.94
17EA ST6	Nwt	90.5	90.12	45FA ST11	Rs	−75.63	−74.79
18VA ST6	Nwt	81.69	80.96	*46KA ST11	Nwt	−65.31	−70.59
*19LA ST6	Rs	82.92	88.77	47KA ST11	Rs	−65.31	−64.42
20DA ST6	Nwt	90.5	90.12	48EA ST11	Rs	−63.03	−62.15
21LA ST11	Rs	−77.2	−76.42	49GA ST11	Rs	−81.17	−80.41
22VA ST11	Rs	−78.6	−77.76	50RA ST11	Rs	−67.41	−66.52
*23RA ST11	Nwt	−67.4	−72.8	51IA ST11	Rs	−77.67	−76.85
*24KA ST11	Nwt	−65.3	−70.59	*52GA ST11	Nwt	−81.17	−85.82
25VA ST6	Nwt	81.69	80.96	*ARC ST11	Nwt	−79.53	81.3
27EA ST6	Nwt	90.5	90.12	**ARC ST6	Nwt	80.71	−62.15
*28EA ST11	Nwt	−63	−68.18				

[a]Arc repressor mutant code: position of mutation, specific mutation, and terminal peptide e.g., 03GA ST6refers 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 Rs points to reduced stability with respect to wild-type Arc repressor.
[c]Resulting Differential (%), e.g., P% = [P(Nwt) − P(Rs)] × 100, this value is positive for mutants predicted as stable and negative for those predicted as unstable.
[d]Analogously, Pcv% = [P_{LOO}(Nwt) − P_{LOO}(Rs)] × 100 considering Leave-one-out (LOO) instead of training probabilities.
*Misclassified compound.
**LOO misclassified compound that was correctly classified in training series.

showing the high stability of the model (both values, cross validation error and percentage of good classification are nearly constant as n increases).(Fig. 3.)

DISCUSSION

As can be seen from Equation (4), the parameter $\Delta\ \Theta_0(Mut - Arc)$ is the only variable that correlates significantly with protein stability. No other variable enters the forward stepwise analysis on LDA. This parameter is the change in entropy for a mutant with respect to the Arc (considered in the standard state). The $\Delta\ \Theta_0(Mut - Arc)$ values are negative (see Table I) for almost all alanine Arc suppressor mutants (except glycine). Thus, the greater the change in charge entropy the higher will be the stabilization of the mutant with respect to the wild-type Arc due to the positive coefficient (92.47) of $\Delta\ \Theta_0(Mut - Arc)$ in the model.

In order to compare the reliability of our methodology, several parameters related to thermodynamic and steric factors were calculated for each mutant. More specifically, surface area[2] and volume were used as steric parameters

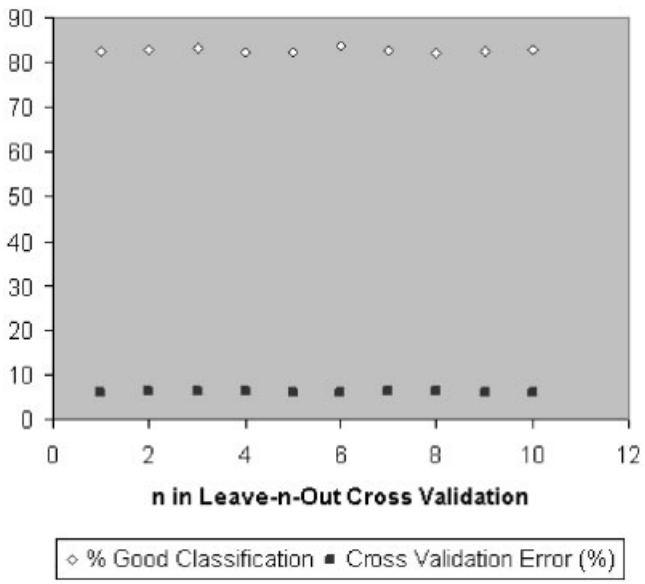

Fig. 3. Leave-n-out cross validation procedure results.

TABLE III. Results of the Comparative Study of the Present Approach With Respect to the Other Five Stability Scoring Functions

Parameter[a]	$\Delta\Theta_0$	D-Fire	Surface	Volume	Log P	Refractivity
%T	81.1	76.9	70.7	62.3	59.0	60.0
%Nwt	71.4	92.9	63.6	53.6	80.8	77.3
%RS	92.0	58.3	78.9	72.0	15.4	38.9
%NC	0.0	3.8	22.6	0.0	26.4	24.5
$\%T_{L\text{-}25\%\text{-}O}$	79.5	71.8	61.5	56.4	48.7	61.5
N	53	53	53	53	53	53
λ	0.56	0.79	0.85	0.92	0.99	0.97
F	39.05	13.9	8.8	4.2	0.5	1.8
p	0.0	0.0	0.0	0.0	0.5	0.2
C	0.643	0.552	0.428	0.259	−0.047	0.175

[a]Parameters verifying model quality: %T, %Nwt, %RS, %NC, $\%T_{L\text{-}25\%\text{-}O}$ are the Total, Near wild-type group, Reduced stability group, Non-classified and total after leave-25%-out Percentages of good classification.

TABLE IV. Results of Forward Stepwise Analysis Considering All Parameters

Model parameters for Forward stepwise analysis with all variables					
Step	Variables	%T	%Nwt	%RS	λ
1	$\Delta\Theta_0$	81.8	70.4	92.0	0.574
2	$\Delta\Theta_0$, D-Fire	82.7	70.0	96.0	0.464
3	$\Delta\Theta_0$, D-Fire, Surface	82.7	77.8	88.0	0.437

Model parameters					
Step	Variables	F	p	F*	p*
1	$\Delta\Theta_0$	37.07	0.00	37.07	0.00
2	$\Delta\Theta_0$, D-Fire	28.32	0.00	11.60	0.00
3	$\Delta\Theta_0$, D-Fire, Surface	20.56	0.00	2.89	0.09

[a]Parameters verifying model quality: %T, %Nwt, %RS, are the Total, Near wild-type group, and Reduced stability group, Percentages of good classification.
*Last entered variable parameters.

and logP and refractivity as thermodynamic ones. In addition, the D-Fire potential was also used as a comparison with a well-known method for predicting protein stability.[49] Table III depicts the results of an LDA to predict mutation-induced stability in this protein using these parameters. As can be seen, several parameters predicted the group of near wild-type stability (Nwt) better than our methodology, particularly D-Fire potential and thermodynamic parameters. On the other hand, our method describes in a very accurate way the reduced stability group (RS) compared with the rest. In general, the MARCH-INSIDE methodology finds the model with the best total percentage of classification from all the methods tested, as well as the model with the highest statistical quality [shown by the values of Wilk's statistic (λ), Fischer Ratio (F), Mathew's coefficient (C), and significance level (p). This method also shows the highest predictability, as demonstrated by its good percentages after the leave-one-out procedure, its stability in the leave-n-out cross validation procedure (Fig. 3 and Table II) and by the total percentage of good classification after leaving 25% of the data out ($\%T_{\text{leave-25\%-out}}$ in Table III).

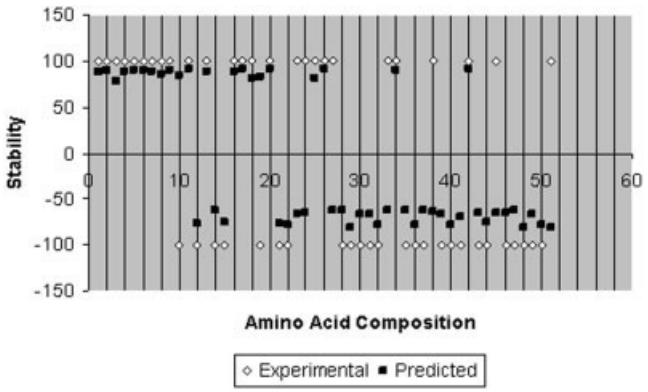

Fig. 4. Comparison between experimental and observed stability versus amino acid position in the protein.

In an attempt to find a better equation for modeling protein stability, we combined all parameters in a single discriminant analysis. The results are shown in the last part of Table IV. Forward Stepwise selects only three parameters when used as a strategy for variable selection.

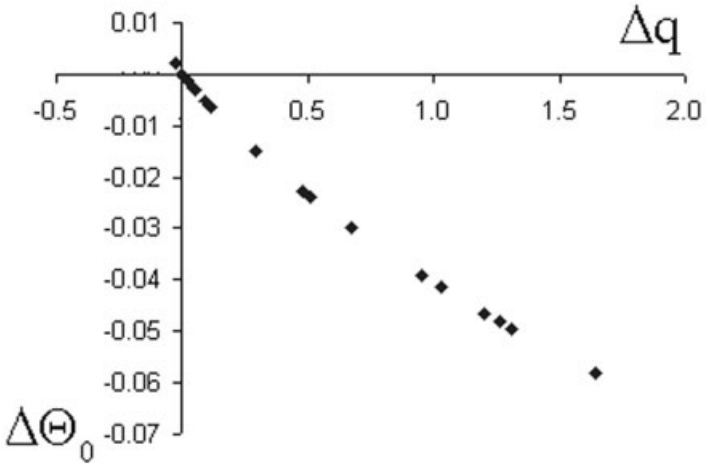

Fig. 5. Entropy change with respect to wild-type Arc repressor by alanine point mutation versus the muted amino acid residue's electronic charge index difference from Equation (5).

Remarkably, $\Delta\ \Theta_0(Mut\ -\ Arc)$ has the higher statistical significance and is selected in the first step. The other two more significant variables were D-Fire potential and Surface Area. Nevertheless, no significant improvement is obtained in the overall fitness after the introduction of these variables.

Figure 4 shows a comparison of experimental and predicted stabilities. These values were obtained by plotting the resulting probability differences (P% in Table II) versus amino acid position in the case of predicted stability and giving a value of 100 to experimentally determined stable mutants and −100 to unstable ones. This figure gives a clearer graphic illustration of the accuracy of this methodology and demonstrates that our method coincides with the experimental determination of non-stability-sensitive mutation areas such as positions 1–8 as well as the stability-sensitive mutation area 28–32.

The fact that this is the only variable that enters into the discriminant function is somewhat puzzling because it is hard to imagine factors that affect stability but do not take into account either connectivity (primary structure) or 3-D structure. However, one must remember that (as depicted in Figure 1) at the initial time (t_0), the connectivity is not considered because the movement of electrons outside its own amino acid residue is not quantified. In this sense, Yuan[32] reported a Markovian model to predict protein subcellular location in a large series of prokaryote and eukaryote organisms. Yuan did not take into account primary or secondary structure and considered only amino acid composition. However, the utility of amino acid composition can be extended by calculating coupling numbers, which consider not only the effect of amino acid composition but also a considerable amount of the sequence-order effects.[32]

In any case, $\Delta\ \Theta_0$ certainly considers something more than amino acid composition, charge information (entropy). It has been recognized that charge is closely related to solvent accessibility and thus to entropy. Based on the mathematical developments and approximations outlined in the methods section, we can easily obtain the following expression for the relationship between $\Delta\ \Theta_0(Mut\ -\ Arc)$ and Δ q (ECI) (the difference between ECI values of aa_i for each mutant and alanine):

$$\Delta\Theta_0(Mut\ -\ Arc)\ =\ \frac{0.05}{\Delta q\ -\ 46.3}\ln\left(\frac{0.05}{46.3\ -\ \Delta q}\right) + \frac{\Delta q\ +\ 0.05}{46.3}ln\left(\frac{\Delta q\ +\ 0.05}{46.3}\right) \quad (5)$$

A plot of Equation (5) over a range of biological Δ q values (ECI) given in Table I is shown in Figure 5.

As stated at the beginning of this section, any factor that decreases the value of $\Delta\ \Theta_0(Mut - Arc)$ would decrease the protein stability. For the relation shown in Figure 5 a linear-like relationship with negative slope is found. We can conclude that the greater the difference in polarity between alanine and aa_i, the greater is the decrease in entropy and thus the stability of the protein is affected. This fact can be explained if one remembers that only covalent and hydrogen-bonding interactions were taken into account to develop the discriminant function. If a highly polar amino acid (i.e., an amino acid that can form polar interactions within a monomer as well between monomers) is substituted by alanine, all these interactions could no longer be maintained. Such a perturbation would introduce significant changes in the electronic distribution $\Delta\ \Theta_0(Mut - Arc)$ and these would affect the stability of the protein.

CONCLUDING REMARKS

In this paper, the MARCH-INSIDE methodology has been extended to protein stability studies. A linear stability/structure relationship has been derived for a series of 53 Arc repressor alanine mutants. This model, fitted by means of LDA, has shown how the entropy of an electro-

static charge distribution may affect protein stability. In this sense, the present work opens the door to the use of simple Markov chain molecular structure descriptors in research in the field of protein functions and properties.

ACKNOWLEDGMENTS

The authors express their gratitude to Dr. Jose Luis Garcia and the Cuban Ministry of Higher Education for financial support and kindness. D. H. González thanks the University of Santiago de Compostela in Spain for kind hospitality and the Xunta de Galicia (PR405A2001/65-0) for partial financial support. This author is also especially indebted to Professor L.B. Kier, USA, for kind peer review of many of the ideas used here.

REFERENCES

1. EUFEPS announcement: conference on optimizing biotech medicines: rational development of therapeutic proteins. Eur J Pharm Sci 2002;15:101–102.
2. Zhou H, Zhou Y. Stability scale and atomic solvation parameters extracted from 1023 mutation experiment. Proteins 2002;49:483–492.
3. Alber T. Mutational effects on protein stability. Annu Rev Biochem 1989;58:765–798.
4. Dill KA, Shortle D. Denatured state of proteins. Annu Rev Biochem 1991;60:795–825.
5. Milla ME, Brown MB, Sauer RT. Protein stability effects of a complete set of alanine substitutions in Arc repressor. Struct Biol 1994;1:518–523.
6. Kowalski RB, Wold S. Pattern recognition in chemistry. In: Krishnaiah PR, Kanal LN. Handbook of statistics. Amsterdam: North Holland Publishing Company; 1982. p 673–697.
7. Estrada E, González H. What are the limits of applicability for graph theoretic descriptor in QSPR/QSAR? Modeling dipole moments of aromatic compounds with Tops-Mode descriptors. J Chem Inf Comp Sci 2003;43:75–84.
8. Estrada E, Peña A. In silico studies for the rational discovery of anticonvulsant compounds. Bioorg Med Chem 2000;8:2755–2770.
9. Golbraik A, Bonchev D, Tropsha A. Novel chirality descriptors derive from molecular topology. J Chem Inf Comput Sci 2001;41: 147–158.
10. Estrada E, Molina E, Uriarte E. Quantitative structure-toxicity relationships using Tops-Mode. 1. Nitrobenzene toxicity to tetrahymena pyriformis. SAR and QSAR Env Res 2001;12:309–324.
11. Kubinyi H, Taylor J, Ramdsen C. Quantitative drug design. In: Hansch C, editor. Comprehensive medicinal chemistry. Vol 4. New York: Pergamon; 1990. p 589–643.
12. Estrada E, Uriarte E. Recent advances on the role of topological indices in drug design discovery research. Curr Med Chem 2001;8:1573–1588.
13. Todeschini R, Consonni V. Handbook of molecular descriptors. Weinheim, Germany: Wiley VCH, 2000.
14. Mansfield ML, Covell DG. A new class of molecular shape descriptors. 1. Theory and properties. J Chem Inf Comput Sci 2002;42:259–273.
15. Shannon CE. The mathematical theory of communication. Urbana, IL: University of Illinois Press, 1955.
16. Bonchev D, Trinjastic N. Information theory, distance matrix and molecular branching. J Chem 1977;67:4517–4533.
17. Kier LB. Use of molecular negentropy to encode structure governing biological activity. J Pharm Sci 1980;69:807.
18. Wilson E.D. The diversity of life. Cambridge, MA: Harvard University Press; 1992.
19. Graham DJ. Information and organic molecules: structure considerations via integer statistics. J Chem Inf Comput Sci 2002;42:215–221.
20. González DH, Olazábal E, Castañedo N, Hernádez SI, Morales A, Serrano HS, González J, Ramos de Armas R. Markovian chemicals "in silico" design (MARCH-INSIDE), a promising approach for computer aided molecular design. II: Experimental and theoretical assessment of a novel method for virtual screening of fasciolicides. J Mol Mod 2002;8:237–245.
21. Gonzáles DH, Gia O, Uriarte E, Hernádez I, Ramos R, Chaviano M, Seijo S, Castillo JA, Morales L, Santana L, Akpaloo D, Molina E, Cruz M, Torres LA, Cabrera MA. Markovian chemicals "in silico" design (MARCH-INSIDE), a promising approach for computer-aided molecular design. I: Discovery of anticancer compounds. J Mol Mod 2003;9:395–407.
22. González DH, Ramos de A R, Molina R. Vibrational markovian modelling of footprints after the interaction of antibiotics with the packaging region of HIV type 1. Bull Math Biol 2003;65:991–1002.
23. González DH, Ramos de A R, Molina R. Markovian negentropies in bioinformatics. 1. A picture of footprints after the interaction of the HIV-1 Ψ-RNA packaging region with drugs. Bioinformatics 2003;19:2079–2087.
24. González DH, Ramos de A R, Uriarte E. In silico markovian bioinformatics for predicting ^{1}Hα-NMR chemical shifts in mouse epidermic growth factor (m-EGF). Online J Bioinf 2002;1:83–95.
25. Ramos de A R, Gonźlez DH, Duran A, Perez M. Stochastic-based descriptors for modeling biological properties of peptides: modeling Angiotensin-Converting Enzyme inhibition of dipeptides. 7th Electronic Conference of Synthetic Organic Chemistry: ECSOC-7, 2003; Proceedings, www.mdpi.net/ecsoc-7/index.htm, c006.
26. González DH, Hernández SI, Uriarte E, Santana L. Symmetry considerations in markovian chemicals "in silico" design (MARCH-INSIDE). I: central chirality codification, classification of ACE inhibitors and prediction of (-receptor antagonist activities. Comput Biol Chem 2003;27:217–227.
27. González DH, Marrero Y, Hernández I, Bastida I, Tenorio I, Nasco O, Uriarte E, Castañedo N, Cabrera M, Aguila E, Marrero O, Morales A, Pérez M. 3D-MEDNES: an alternative "in silico" technique for chemical research in toxicology. 1. Prediction of chemically induced agranulocytosis. Chem Res Tox 2003;16:1318–1327.
28. Borodovsky M, Koonin EV, Rudd KE. New genes in old sequence: a strategy for finding genes in the bacterial genome. Trends Biochem Sci 1994;19:309–313.
29. Borodovsky M, Macininch JD, Koonin EV, Rudd KE, Médigue C, Danchin A. Detection of new genes in a bacterial genome using Markov Models for three gene classes. Nucleic Acid Res 1995;23: 3554–3562.
30. Hughey R, Krogh A. Hidden Markov models for sequence analysis: extension and analysis of the basic method. CABIOS 1996;12:95–107.
31. Chou KC. Prediction and classification of α-turn types. Biopolymers 1997;42:837–853.
32. Yuan Z. Prediction of proteins subcellular location using Markov chain models. FEBS Lett 1999;451:23–26.
33. Hua S, Sun Z. Support vector machine approach for protein subcellular localization prediction. Bioinformatics 2001;17:721–728.
34. Hubbard TJ, Park J. Fold recognition and ab initio structure predictions using hidden Markov models and β-strand pair potential. Proteins 1995;23:398–402.
35. Krogh A, Brown M, Mian IS, Sjeander K, Haussler D. Hidden Markov Models in computational biology: applications to protein modeling. J Mol Biol 1994;235:1501–1531.
36. Di Francesco V, Munson PJ, Garnier J. Foresst: fold recognition from secondary structure predictions of proteins. Bioinformatics 1999;15:131–140.
37. Chou KC. Prediction of protein signal sequences. Curr Protein Pept Sci 2002;3:615–622.
38. Chou KC. Prediction of signal peptides using scaled window. Peptides 2001:22:1973–1979.
39. Chou KC. Review: Prediction of tight turns and their types in proteins. Analytical Biochemistry 2000;286:1–16.
40. Chou KC. A vectorized sequence-coupling model for predicting HIV protease cleavage sites in proteins. J Biol Chem 1993;268: 16938–16948.
41. Chou KC. Review: Prediction of HIV protease cleavage sites in proteins. Anal Biochem 1996;233:1–14.
42. Chou KC, Zhang CT. Studies on the specificity of HIV protease: an

application of Markov chain theory. J Protein Chem 1993;12:709–724.
43. Freund JA, Poschel T. Stochastic processes in physics, chemistry, and biology. In: Lecture Notes Phys. Berlin, Germany: Springer-Verlag; 2000.
44. Collantes ER, Dunn WJ. Amino acid chain descriptors for QSAR studies of peptides analogues. J Med Chem 1995;38:2705–2713.
45. Bharucha-Reid AT. Elements of theory of Markov Process on the application, McGraw-Hill Series in Probability and Statistic. New York: McGraw-Hill; 1960. p 167–434.
46. Gnedenko B. The theory of probability. Moscow: Mir Publishers; 1978. p 107–112.
47. Hernández I, González DH. MARCH-INSIDE version 2.0, 2002 (Markovian Chemicals "In Silico" Design), Chemicals Bio-actives Center, Central University of "Las Villas," Cuba. This is a preliminary experimental version; a future professional version will be available to the public. For further information about it, e-mail the corresponding author, humbertogd@vodafone.es
48. Hypercube, Inc. Hyperchem, 2002: 7.0.
49. Zhou H, Zhou Y. Distance-scaled, finite ideal gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. Protein Sci 2002;11:2714–2726. This service is available at http://theory.med.buffalo.edu/.
50. Bernstin FC, Koetzle TF, Wiliams GJB. The Protein Data Bank: A computer-based archival file for macromolecular structures. J Mol Biol 1977;112:535–542.
51. Wang R, Liu L, Lai L, Tang Y. Score: a new empirical method for estimating the binding affinity of a protein-ligand complex. J Mol Model 1998;4:379–394.
52. Chou KC, Zhang CT. Prediction of protein structural classes. Crit Rev Biochem Mol Biol 1995;30:275–349.
53. Zhou GP. An intriguing controversy over protein structural class prediction. J Protein Chem 1998;17:729–738.
54. Zhou GP, Assa-Munt N. Some insights into protein structural class prediction. Proteins 2001;44:57–59.