Available online at www.sciencedirect.com
SCIENCE DIRECT®
Bioorganic & Medicinal Chemistry 13 (2005) 3641–3647
Bioorganic & Medicinal Chemistry

Proteins Markovian 3D-QSAR with spherically-truncated average electrostatic potentials

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

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

Received 18 December 2004; revised 16 March 2005; accepted 21 March 2005
Available online 20 April 2005

Abstract—Proteins 3D-QSAR is an emerging field of bioorganic chemistry. However, the large dimensions of the structures to be handled may become a bottleneck to scaling up classic QSAR problems for proteins. In this sense, truncation approach could be used as in molecular dynamic to perform timely calculations. The spherical truncation of electrostatic field with different functions breaks down long-range interactions at a given cutoff distance (r_{off}) resulting in short-range ones. Consequently, a Markov chain model may approach to the average electrostatic potentials of spatial distribution of charges within the protein backbone. These average electrostatic potentials can be used to predict proteins properties. Herein, we explore the effect of abrupt, shifting, force shifting, and switching truncation functions on 3D-QSAR models classifying 26 proteins with different functions: lysozymes, dihydrofolate reductases, and alcohol dehydrogenases. Almost all methods have shown overall accuracies higher than 73%. The present result points to an acceptable robustness of the MC for different truncation schemes and r_{off} values. The results of best accuracy 92% with abrupt truncation coincide with our recent communication. We also developed models with the same accuracy value for other truncation functions; however they are more complex functions. PCA analysis for 152 non-homologous proteins has shown that there are five main eigenvalues, which explain more than 87% of the variance of the studied properties. The present molecular descriptors may encode structural information not totally accounted for the previous ones, so success with these descriptors could be expected when classic fails. The present result confirms the utility of our Markov models combined with truncation approach to generate bioorganic structure protein molecular descriptors for QSAR.

1. Introduction

Markov chains (MC) models have been used for analyzing biological sequence data and to find new genes from the open reading frames. Another use is data-based searching and multiple sequence alignment of protein families and protein domains. Hubbard and Park used amino acid sequence-based hidden MC models to predict secondary structures. In this sense, Krogh et al.[4] have also proposed a hidden MC model architecture. In addition, Markov's stochastic process has been used for protein folding recognition. This approach can also be used for the prediction of protein signal sequences[5,6] applied MC models to predict beta turns and their types, and the prediction of protein cleavage sites by HIV protease. Anyhow, many works have not been reported on Markov models for the generation of molecular descriptors encoding proteins 3D structure facing QSAR (quantitative structure–activity relationships).

Proteins are highly charged molecules for which an accurate treatment of the long-range electrostatic interactions is very important in molecular dynamics (MD) approximations to proteins structure.[7] Non-bonded pairwise interactions between atoms or groups are usually truncated at a specific cutoff distance (r_{off}) to reduce the number of interactions and thereby the required computational time for the simulation. The interaction energy or force can be *truncated* abruptly at the cutoff distance, or some kind of smoothing scheme can be applied, either on the whole range 0, r, r_{off} (a *shift*), or over a limited region r_{on}, r, r_{off} (a *switch*). The search for theoretic approaches reaching to new molecular descriptors for biopolymers has begun more after in spite of an early (pioneer) work of Flory[8] on the radius of gyration. More recently, other approaches appeared, which are

Keywords: QSAR; Markov models; Proteins function; Long-range interaction; Electrostatic potential; Linear discriminant analysis; Principal components analysis.
* Corresponding author. Address: Department of Organic Chemistry, Faculty of Pharmacy, University of Santiago de Compostela, 15706, Spain. Tel.: +34 981563100x14938; fax: +34 981594912; e-mail: humbertogd@vodafone.es

0968-0896/$ - see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.bmc.2005.03.041

potential sources, define, or apply in some extent biopolymer descriptors, such as the Arteca's mean over crossing number,[9] the Randič's band average widths,[10–12] Emini surface index, the SDA (sum of cosines of dihedral angles), Kyle–Dolittle hydrophobicity,[13] and the I_3 index[14] among others. In any case, the search of molecular descriptors for biopolymers structure facing QSAR studies is an emerging area.

Our group has reported some interesting MC models to generate molecular descriptors encoding biopolymers structure[3,15,1,2,16,17] but ignoring truncation approaches. Consequently, we will describe herein a number of studies that have focused on the advantages or disadvantages of different truncation methods for long-range electrostatic interactions on proteins 3D-QSAR using MC molecular descriptors.

2. Methods

Consider a representation for a protein, described as a static model, which considers a spatial distribution of aminoacids, with 3D coordinates (x_i, y_i, z_i) coinciding with those for a reference atom in the aminoacid specifically the Cα. In this case, every pair of aminoacid in the protein backbone (i,j) presents a pairwise electrostatic interaction with energy E_{ij}. The electrostatic charge (q_i) will be considered to be equal to the electronic charge of the aminoacid reported by Collantes and Dunn.[18] All of these E_{ij} may be determined using Coulomb's formula. If we then arrange all these interaction energies in a matrix and normalize the values dividing by row sums, we obtain a stochastic matrix ${}^1\Pi(x,y,z,q)$. This step makes it possible to study the propagation of the electrostatic interactions within the protein backbone as a MC. In doing so, the elements of ${}^1\Pi(x,y,z,q)$ may be considered as the probabilities $({}^1p_{ij})$ with which the aminoacid i presents a truncated electrostatic interaction of energy E_{ij}, with the aminoacid j placed at a distance r_{ij}.[18]

$$
\begin{aligned}
{}^1p_{ij} &= \frac{S_{ij}^2(r)\cdot E_{ij}}{\sum_{m=1}^{\delta+1} S_{im}^2(r)\cdot E_{im}} = \frac{S_{ij}^2(r)\cdot q_i\cdot q_j/r_{ij}^2}{\sum_{k=1}^{\delta+1} S_{im}^2(r)\cdot q_i\cdot q_m/r_{im}^2} \\
&= \frac{S_{ij}^2(r)\cdot q_i\cdot q_j/r_{ij}^2}{q_i\cdot\sum_{k=1}^{\delta+1} S_{im}^2(r)\cdot q_m/r_{im}^2}
\end{aligned}
\tag{1a}
$$

As can be seen, it is equivalent to use energy (Eq. 1a) or an electrostatic potential (φ_j) interpretation (Eq. 1b)

$$
{}^1p_{ij} = \frac{S_{ij}^2(r)\cdot q_j/r_{ij}^2}{\sum_{m=1}^{\delta+1} S_{im}^2(r)\cdot q_m/r_{im}^2} = \frac{S_{ij}^2(r)\cdot \varphi_{ij}}{\sum_{m=1}^{\delta+1} S_{im}^2(r)\cdot \varphi_{im}} \tag{1b}
$$

In Eqs. 1a and 1b, the sum considers all the δ aminoacids that have a spherical truncated interaction with the aminoacid i. $S_{ij}(r)$ is the truncation function, which has different formulation in dependence of the method used, that is, a shifting function like in Eq. 2a or a force-shifting function like in Eq. 2b. Alternatively, we will explore also a switching function like in Eq. 2c:[19]

$$
S_{ij}(r) = \begin{cases} \left(1-\left(\dfrac{r}{r_{\text{off}}}\right)^2\right)^2, & r \leqslant r_{\text{off}} \\ 0, & r > r_{\text{off}} \end{cases} \tag{2a}
$$

$$
S_{ij}(r) = \begin{cases} \left(1-\dfrac{r}{r_{\text{off}}}\right)^2, & r \leqslant r_{\text{off}} \\ 0, & r > r_{\text{off}} \end{cases} \tag{2b}
$$

$$
S_{ij}(r) = \begin{cases} 1, & r \leqslant r_{\text{on}} \\ \dfrac{(r_{\text{off}}^2 - r^2)^2\cdot(r_{\text{off}}^2 + 2r^2 - 3r_{\text{on}}^2)}{(r_{\text{off}}^2 - r_{\text{on}}^2)^3}, & r_{\text{on}} < r < r_{\text{off}} \\ 0, & r \geqslant r_{\text{off}} \end{cases} \tag{2c}
$$

Due to truncation restrictions, the aminoacid a_0 is only able to interact with the aminoacid a_1 by means of direct interaction. The parameter of the MC is the topologic distance or number of steps (k) one interaction needs to propagate from one aminoacid to other instead of the time, which is the more classic MC parameter. However, one should note that this number of elemental steps (k) one truncated interaction uses to propagate throughout space is given at corresponding discrete time intervals $(\Delta t_k = t_{k+1} - t_k = k)$ like in almost all MC applications.[15]

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 ${}^0\boldsymbol{\varphi}$. It can be supposed that, after this initial situation, all the amino acids interact with electrostatic energy ${}^1E_{ij}$ with every other aa_j in the protein.

By using MC theory, it is possible to develop a simple 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.[2]

$$
\begin{aligned}
\xi_k &= \sum_{j=1}^{n} {}^{\text{A}}p_k(j)\cdot\varphi_j = {}^0\boldsymbol{\pi}^{\mathbf{T}}\cdot{}^k\Pi\cdot{}^0\boldsymbol{\varphi} \\
&= {}^0\boldsymbol{\pi}^{\mathbf{T}}\cdot({}^1\Pi)^k\cdot{}^0\boldsymbol{\varphi}
\end{aligned}
\tag{3}
$$

It is remarkable that the average electrostatic potentials ξ_k depend on the absolute probabilities ${}^{\text{A}}p_k(j)$ with which the amino acids interact with other amino acids placed at distance k. The potential ξ_k also depends on the initial unperturbed electrostatic potential of the aminoacid. In matrix form represented above, the ${}^{\text{A}}p_k(j)$ are

calculated with the vector ${}^{0}\boldsymbol{\pi}$, of absolute initial probabilities, and the matrix ${}^{1}\Pi$ using the Chapman–Kolgomorov equations.[16]

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 as follows this expansion for the tripeptide AVW (Ala-Val-Trp):

$$\xi_k = \left[{}^{A}p_0(A), {}^{A}p_0(V), {}^{A}p_0(W)\right] \cdot \begin{bmatrix} {}^{1}p_{AA} & {}^{1}p_{AV} & 0 \\ {}^{1}p_{VA} & {}^{1}p_{VV} & {}^{1}p_{VW} \\ 0 & {}^{1}p_{WV} & {}^{1}p_{WW} \end{bmatrix} \cdot \begin{bmatrix} \varphi_A \\ \varphi_V \\ \varphi_W \end{bmatrix} = {}^{A}p_k(A) \cdot \varphi_A + {}^{A}p_k(V) \cdot \varphi_V + {}^{A}p_k(W) \cdot \varphi_W \quad (4)$$

All calculations were carried out with our software BIOMARKS® *version 1.0.* (BIOinformatics MARKovian Studio).[16]

3. Results and discussion

In order to determine the effect of using different long-range electrostatic field truncation approaches in polymers 3D-QSAR, we have developed a linear discriminant analysis to find a QSAR for 26 proteins. These proteins in spite of similar folding have three different biological activities, namely: Lysozymes (L), dihydrofolate reductases (DR), and alcohol dehydrogenases (AD). Briefly, all the truncation methods were used for seeking significant QSAR models.

(a) Abrupt truncation function: This approach presented 92% higher accuracy at $r_{off} = 50\%$ of the Van der Waal distance among atoms, in coincidence with our previous reports[2] and models

Table 1. Accuracies for the different truncation functions

r_{off} (%)	Scheme	AD (%)	L (%)	DR (%)	Total (%)	λ	F	R_c
Truncation scheme: abrupt truncation								
50		**89**	**100**	**86**	**92**	**0.29**	**5.99**	**0.82**
60		**89**	**100**	**86**	**92**	**0.25**	**6.86**	**0.80**
70		89	100	71	88	0.26	6.80	0.79
80		89	100	57	85	0.26	6.80	0.79
90		89	100	71	88	0.26	6.80	0.79
Truncation scheme: force-shifting function								
7		100	70	43	73	0.27	6.55	0.80
8	AFSHS	100	70	57	77	0.27	6.51	0.80
9		100	70	57	77	0.27	6.49	0.80
10		100	70	57	77	0.27	6.47	0.80
11		**89**	**100**	**86**	**92**	**0.27**	**6.45**	**0.79**
12	**AFSH**	**89**	**100**	**86**	**92**	**0.27**	**6.51**	**0.79**
13		**89**	**100**	**86**	**92**	**0.27**	**6.59**	**0.79**
18	**AFSHL**	**89**	**100**	**86**	**92**	**0.25**	**7.06**	**0.80**
Truncation scheme: shifting function								
7		33	50	43	42	0.61	1.97	0.55
8		33	50	43	42	0.61	1.98	0.55
9		100	70	57	77	0.26	6.62	0.80
10		100	70	43	73	0.27	6.58	0.80
11		100	70	43	73	0.27	6.54	0.80
12	ASH	89	70	57	73	0.27	6.53	0.80
13		**89**	**100**	**86**	**92**	**0.26**	**6.67**	**0.79**
Truncation scheme: switching function with $r_{off} = 12$ and variable r_{on}								
5		89	70	43	69	0.26	6.67	0.80
6		**89**	**100**	**86**	**92**	**0.26**	**6.70**	**0.79**
7		**89**	**100**	**86**	**92**	**0.26**	**6.74**	**0.79**
8	ASW	89	90	86	88	0.13	12.53	0.87
9		**89**	**100**	**86**	**92**	**0.25**	**6.90**	**0.79**
10		**89**	**100**	**86**	**92**	**0.25**	**6.90**	**0.79**
11		**89**	**100**	**86**	**92**	**0.25**	**6.95**	**0.79**
Truncation scheme: switching function with $r_{on} = 8$ and variable r_{off}								
9		89	70	43	69	0.26	6.71	0.80
10		78	70	43	65	0.26	6.62	0.79
11		89	70	43	69	0.26	6.65	0.80
12	ASW	**89**	**100**	**86**	**92**	**0.26**	**6.79**	**0.79**
13		**89**	**100**	**86**	**92**	**0.25**	**6.91**	**0.79**
14		**89**	**100**	**86**	**92**	**0.25**	**6.99**	**0.79**
15		**89**	**100**	**86**	**92**	**0.25**	**7.03**	**0.79**

implemented in almost all docking and MD softwares.[20] In addition, at $r_{off} = 60\%$ the higher accuracy of 92% is presented too. However, the method of accuracy abruptly decays to 88% and less for every $r_{off} > 60\%$, see Table 1 and Figure 1.

(b) Force-shifting function: The behavior of the accuracy in this approach has a tendency to grow. For this truncation scheme, the changes in overall accuracy 73–77%, 77–92% have taken place at $r_{off} = 7$, $r_{off} = 10$, from $r_{on} = 11$ the maximum value of accuracy stays constant. It is notable that the atom-based approach with the force-shift method[21] and a short cutoff of 8.0 Å (AFSHS)[22] presented less accuracy than its middle and longer range analogues AFSH and AFSHL, see Table 1 and Figure 1.

(c) Shifting function: The QSAR models with r_{off} equal to 7 or 8 presented low values of the canonical regression coefficient $R_c = 0.55$ and the most low values of overall accuracy in comparison with the other truncation functions (42%). At $r_{off} = 13$ Å, the model has the highest value.

(d) Switching function: We developed two experiments expanding the domain of the switching function both inward ($r_{off} = 12$ and r_{on} variable) and outward ($r_{on} = 8$ and r_{off} variable). The graphic for the experiment with $r_{off} = 12$ presented two minimum values of overall accuracy at $r_{on} = 5$ (69%) and $r_{on} = 8$ (88%) the other points remain constant at the most high value of overall accuracy (92%). In this study, there are a few variations. For the second experiment, all overall accuracy stays approximately constant at $r_{off} = 9$–11 Å and from this last point it grows abruptly to the maximum value (92%) to remain constant.

All the QSAR models studied have three equations as a consequence of a three group (L, AD, DR) LDA analysis and three variables. Nevertheless, we are going to report only the better QSAR model found herein taking into consideration the higher accuracy and simpler cutoffs procedure. Table 2 summarizes the classification for all proteins studies with different truncation approaches.

The model derived with abrupt truncation function presented 92% higher accuracy at $r_{off} = 50\%$ and at $r_{off} = 60\%$ (for more information about the equations, see the supplementary material).

$$\begin{aligned} L &= 1545.0 \times \xi_0(3) - 308.4 \times \xi_0(4) + 5.5 \times \xi_5(4) - 49.5 \\ AD &= 1457.2 \times \xi_0(3) - 529.2 \times \xi_0(4) + 6.7 \\ &\quad \times \xi_5(4) - 48.1 \\ DR &= 1606.5 \times \xi_0(3) - 456.0 \times \xi_0(4) + 6.6 \\ &\quad \times \xi_5(4) - 54.4 \end{aligned} \tag{5a}$$

$$\begin{aligned} L &= 1791.9 \times \xi_0(3) - 459.3 \times \xi_0(4) + 7.2 \times \xi_5(4) - 59.1 \\ AD &= 1633.3 \times \xi_0(3) - 635.3 \times \xi_0(4) + 7.9 \\ &\quad \times \xi_5(4) - 52.2 \\ DR &= 1917.5 \times \xi_0(3) - 646.4 \times \xi_0(4) + 8.8 \\ &\quad \times \xi_5(4) - 69.5 \end{aligned} \tag{5b}$$

In the symbol $\xi_k(o)$, k is the topological distance between the aminoacids considered and (o) is the orbit. That is to say, 0, 1, 2, 3, and 4 are specific groups or collections of aminoacids placed at protein core (0), inner (1), middle region (2), surface (3), or in every place (4).[2]

As can be noted, both equations are robust with respect to the contributions signs to the enzymatic action of the molecular descriptors. For instance, in all cases the contribution is positive for $\xi_0(3)$, the initial surface electrostatic potential, having different susceptibilities for each family of enzymes. However, this positive contribution is negatively regulated by an overall negative contribution of the rest of the protein $\xi_0(4)$. It means that the magnitude of the difference on the initial electrostatic potential between the surface and the rest of the protein plays a role in the enzymatic activity for the studied proteins. Finally, the model also recognizes the importance of a positive contribution for the evolution of the overall electrostatic potential for the whole enzyme after the initial interaction $\xi_0(5)$.

Several years ago, Randič introduced the orthogonalization process for molecular indices as a way to improve the statistical interpretation of the model when descriptors are correlated with one another.[23–25] For that reason, we developed a principal component analysis (PCA) using for this study a datum of 152 non-homologous proteins.[26]

After applying Randič's procedure, new orthogonal ${}^{m}O_{ko}$ takes into account only the information related to the movement of the electrons exactly at time k and not at times $<k$. The symbol O means orthogonal and m is the degree of importance of the descriptor to explain the property determined by the order in which it is selected by forward stepwise analysis[27] and o is the orbit.

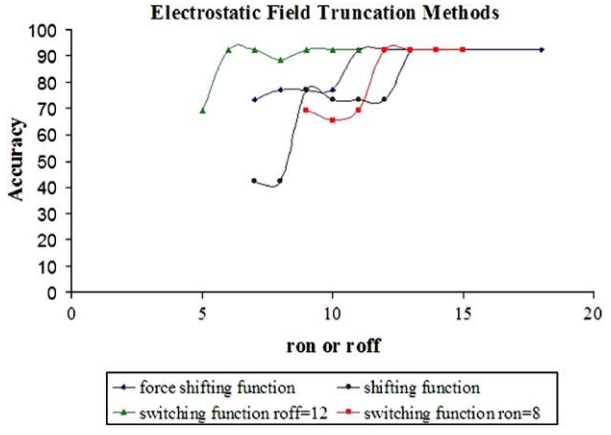

Figure 1. Accuracies for the abrupt truncation function.

Table 2. Classification of proteins according to different truncation schemes

Protein PDB ID	Observed protein family	Classification with different Truncation Approaches							
		Abrupt truncation		Force-shifting function			Shifting function		Switching function
		r_{off} = 50	r_{off} = 60	AFSHS r_{off} = 8	AFSH r_{off} = 12	AFSHL r_{off} = 18	ASH r_{off} = 12	r_{off} = 13	ASW r_{off} = 12, r_{on} = 8
4lyt A	L	L	L	L	L	L	L	L	L
1ghl A	L	L	L	L	L	L	L	L	L
1ghl B	L	L	L	DR	L	L	DR	L	L
1hnl	L	L	L	L	L	L	L	L	L
1lmn	L	L	L	L	L	L	L	L	L
1lz1	L	L	L	L	L	L	L	L	L
2eql	L	L	L	DR	L	L	DR	L	L
2ihl	L	L	L	L	L	L	L	L	L
4lyt A	L	L	L	L	L	L	L	L	L
135l	L	L	L	DR	L	L	DR	L	L
2ohx A	AD	AD	AD	AD	AD	AD	AD	AD	AD
2ohx B	AD	AD	AD	AD	AD	AD	AD	AD	AD
1cdd A	AD	AD	AD	AD	AD	AD	AD	AD	AD
1cdd B	AD	AD	AD	AD	AD	AD	AD	AD	AD
1deh A	AD	AD	AD	AD	AD	AD	AD	AD	AD
1deh B	AD	AD	AD	AD	AD	AD	AD	AD	AD
1qor A	AD	AD	AD	AD	AD	AD	AD	AD	AD
1qor B	AD	AD	AD	AD	AD	AD	AD	AD	AD
1uok	AD	DR	DR	AD	DR	DR	DR	DR	DR
8dfr	DR	DR	DR	DR	DR	DR	DR	DR	DR
1ai9 B	DR	AD	AD	AD	AD	AD	AD	AD	AD
1drf	DR	DR	DR	DR	DR	DR	DR	DR	DR
1dyr	DR	DR	DR	DR	DR	DR	DR	DR	DR
4dfr	DR	DR	DR	AD	DR	DR	AD	DR	DR
3dfr	DR	DR	DR	DR	DR	DR	DR	DR	DR
1ai9 A	DR	DR	DR	AD	DR	DR	AD	DR	DR

In Table 3 we show the eigenvalues, the percentage of the variance explained by them individually, and the cumulative percentage of the variance of the physical properties reproduced by successive addition of eigenvalues. As can be seen in this table, there are five main eigenvalues that explain more than 87% of the variance of the studied properties.

The factor loadings of the PCA are shown in Table 4. They were obtained after a Varimax normalized rotation of the factor space. As we can observe in this table, there are two groups of variables that correlate factor 1 (F1) and factor 2 (F2), respectively. The first group contains the variables n, OSP, R_g, and $^6O_{54}$ and the second group contains the variables H%, S%, and I_3. This information about the second group coincides with other results,[14] since these descriptors are physically related; the secondary structure of a protein is directly related to the degree of folding of the main chain. A protein is considered more folded than another if it contains a greater percentage of helix and/or a lesser strand percentage.

In the first group, the variable occluded surface packing, OSP, measures the interatomic occluded surface area for

Table 3. Factor analysis explained variance

Factor	Eigenvalue	% Total variance	Cumulative eigenvalue	Cumulative variance
F1	3.14	31.40	3.14	31.40
F2	2.64	26.42	5.78	57.82
F3	1.01	10.09	6.79	67.91
F4	1.00	10.01	7.79	77.92
F5	0.94	9.44	8.74	87.35

Table 4. Factor loadings for PCA analysis (varimax normalized rotation)

Variables	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5
n[a]	**0.94**	0.01	0.00	0.11	0.09
H%[b]	0.07	**0.96**	0.04	0.00	−0.03
S%[c]	−0.04	**−0.96**	−0.03	0.00	0.05
OSP[d]	**0.73**	0.06	0.08	0.18	0.18
R_g[e]	**0.91**	0.08	0.00	0.11	0.12
I_3[f]	0.04	**0.98**	0.06	−0.02	−0.05
$^1O_{03}$	0.02	0.08	**0.99**	−0.01	0.00
$^1O_{04}$	−0.13	0.02	0.01	**−0.95**	0.05
$^3O_{24}$	0.13	−0.09	−0.01	−0.05	**0.95**
$^6O_{54}$	**−0.73**	−0.04	0.05	0.27	0.26
Expl. var	2.82	2.84	1.00	1.04	1.04
Prp. totl	0.28	0.28	0.10	0.10	0.10

[a] n: number of aminoacids.
[b] H%: percentage of helix.
[c] S%: percentage of strands.
[d] OSP: occluded surface packing.
[e] R_g: the radius of gyration of the protein.
[f] I_3: folding degree index.

each atom in the protein.[28] The radius of gyration of the protein, R_g, has been widely used as a measure of the global compactness and also as a measure of packing,[8] $^{6}O_{54}$ is a global variable correlate with n and the other descriptors. On the other hand, factor 3 (F3), factor 4 (F4), and factor 5 (F5) are related with the variables $^{1}O_{03}$, $^{1}O_{04}$, and $^{3}O_{24}$, respectively.

The property space projected in the three principal components is illustrated in Figure 2 in three different forms, Figure 2a with factor 1 versus factor 2 versus factor 3, Figure 2b with factor 1 versus factor 2 versus factor 4, and Figure 2c with factor 1 versus factor 2 versus factor 5. These results show that the present molecular descriptors may encode structural information not totally accounted for the previous ones, so success with these descriptors could be expected when classic fails. This fact coincides with results reported on modeling proteins function where classic molecular descriptors such as partition coefficient, molecular refractivity, and polarizability presented only 80.8% with respect to 92% for our molecular descriptors in the same data set.[2]

4. Conclusions

In summary, we would draw five main conclusions from this study:

1. This work introduces for the first time the definition of a new stochastic molecular descriptor named Markov average electrostatic potentials $\xi_k(o)$, for proteins QSAR.
2. In order to save the time of calculation for these molecular descriptors, different truncation approaches may be used.
3. Studies should be carried out to determine at which conditions of number of variables, truncation approach, and cutoffs can one find the best QSAR models.
4. In this specific case, we show how the model applies to the study of proteins function with the best results in terms of accuracy and simplicity for abrupt truncation.
5. PCA analysis for 152 non-homologous proteins shows that Markov molecular descriptors may encode structural information not totally accounted

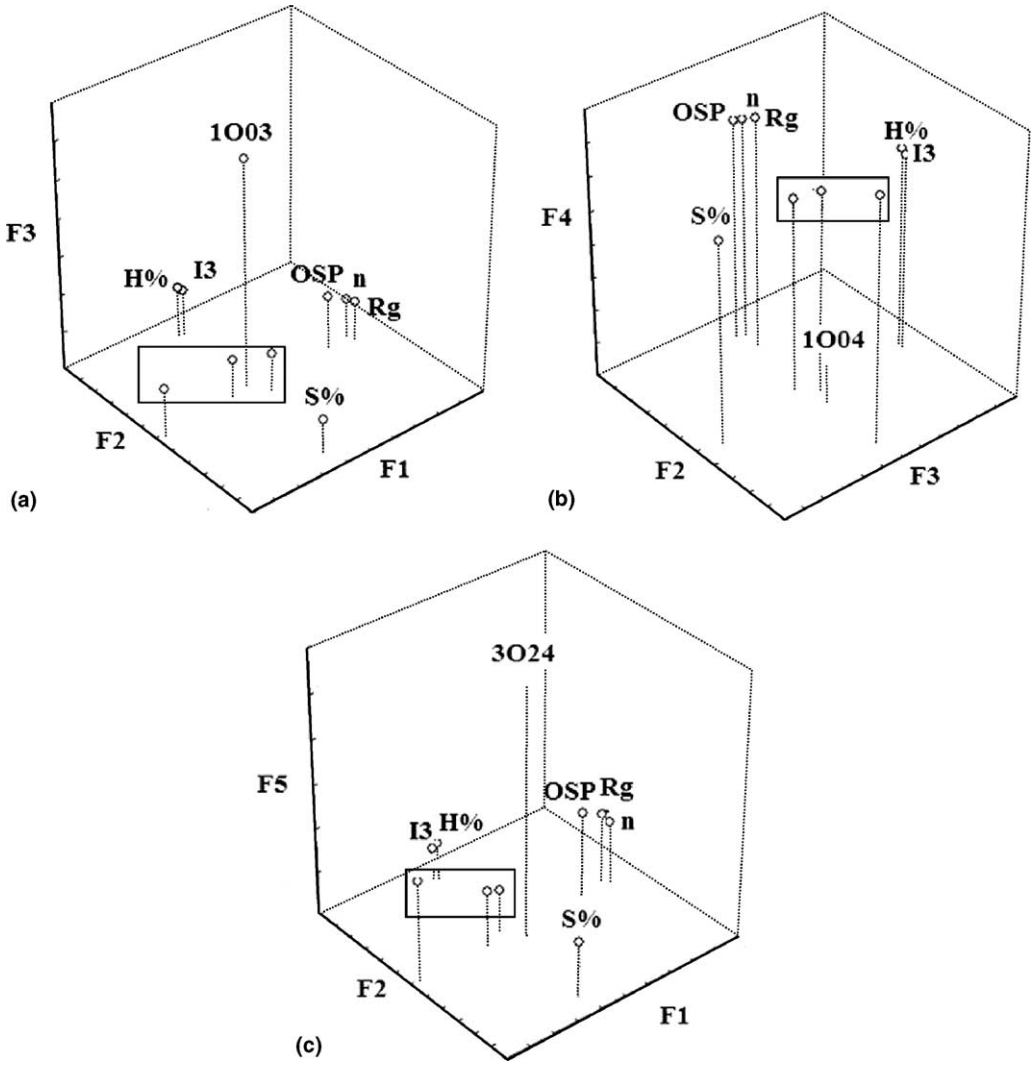

Figure 2. 3D PCA spaces obtained with principal factors 1, 2, and an additional third factor that can be either (a) factor 3, (b) factor 4, or factor (5).

for the previous ones, so success with these descriptors could be expected when classic fails.[29,30]

Acknowledgements

Thanks are given to the Xunta of Galicia for partial financial support (Project BTF20301PR).

Supplementary data

Supplementary data associated with this article can be found in the online version, at doi:10.1016/j.bmc.2005.03.041.

References and notes

1. González-Díaz, H.; Molina, R.; Uriarte, E. *Polymer* **2004**, *45*, 3845.
2. González-Díaz, H.; Molina, R.; Uriarte, E. *Bioorg. Med. Chem. Lett.* **2004**, *14*, 4691.
3. González-Díaz, H.; Ramos, R.; Molina, R. *Bioinformatics* **2003**, *19*, 2079.
4. Krogh, A.; Brown, M.; Mian, I. S.; Sjeander, K.; Haussler, D. *J. Mol. Biol.* **1994**, *235*, 1501.
5. Di Francesco, V.; Munson, P. J.; Garnier, J. *Bioinformatics* **1999**, *15*, 131.
6. Chou, K. C. *Biopolymers* **1997**, *42*, 837.
7. McCammon, J. A.; Harvey, S. C. Cambridge University Press: Cambridge, UK, 1987.
8. Flory, P. J. Cornell University Press: Ithaca, NY, 1953.
9. Arteca, G. A. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 550.
10. Randič, M. In *Encyclopedia of Computational Chemistry*; Schleyer, P. V. R., Ed.; John Wiley & Sons: New York, 1998; Vol. 5, p 3018.
11. Randić, M.; Vračko, M.; Nandy, A.; Basak, S. C. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1235.
12. Randić, M.; Balaban, A. T. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 532.
13. Hua, S.; Sun, Z. *Bioinformatics* **2001**, *17*, 721.
14. Estrada, E. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1238.
15. 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. Model.* **2003**, *9*, 395.
16. González-Díaz, H.; Molina, R.; Hernández, I. BIO-MARKS®, **2004**, *version 1.0*. contact information: humbertogd@vodafone.es or humbertogd@usc.es.
17. González-Díaz, H.; Bastida, I.; Castañedo, N.; Nasco, O.; Olazabal, E.; Morales, A.; Serrano, H. S.; Ramos, R. *Bull. Math. Biol.* **2004**, *66*, 1285.
18. Collantes, E. R.; Dunn, W. J. *J. Med. Chem.* **1995**, *38*, 2705.
19. Norberg, J.; Nilsson, L. *Biophys. J.* **2000**, *79*, 1537.
20. Navarro, E.; Fenude, E.; Celda, B. *Biopolymers* **2002**, *64*, 198.
21. Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. *J. Comput. Chem.* **1983**, *4*, 187.
22. Brooks, C. L., III; Pettitt, B. M.; Karplus, M. *J. Chem. Phys.* **1985**, *83*, 5897.
23. Randič, M. Correlation of enthalpy of octanes with orthogonal connectivity indices. *J. Mol. Struct. (THEOCHEM)* **1991**, *233*, 45.
24. Randič, M. *New J. Chem.* **1991**, *15*, 517.
25. Randič, M. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 311.
26. Fleming, P. J.; Richards, F. M. *J. Mol. Biol.* **2000**, *299*, 487.
27. Estrada, E.; Perdomo, I.; Torres, J. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1561.
28. Pattabiraman, N.; Ward, K. B.; Fleming, P. J. *J. Mol. Recognit.* **1995**, *8*, 334.
29. González-Díaz, H.; Uriarte, E. *Biopolymers* **2005**, *77*, 296.
30. González-Díaz, H.; Uriarte, E.; Ramos de Armas, R. *Bioorg. Med. Chem.* **2005**, *13*, 323.