Available online at www.sciencedirect.com
SCIENCE DIRECT®
Polymer 45 (2004) 3845–3853
polymer
www.elsevier.com/locate/polymer

Stochastic molecular descriptors for polymers. 1. Modelling the properties of icosahedral viruses with 3D-Markovian negentropies

Humberto González Díaz[a,b,*], Reinaldo Molina[a,c], Eugenio Uriarte[b]

[a]*Chemical Bioactives Center, Central University of Las Villas, Santa Clara 54830, Villa Clara, Cuba*
[b]*Department of Organic Chemistry, Faculty of Pharmacy, University of Santiago de Compostela, 15706 Santiago de Compostela, Spain*
[c]*Universität Rostock, FB Chemie, Albert-Einstein-Str. 3a, D 18059 Rostock, Germany*

Received 19 January 2004; received in revised form 23 March 2004; accepted 23 March 2004

Abstract

The application of topologic descriptors to polymer data is considered to be an emerging field. Recently, MARCH-INSIDE was introduced, a topologic approach based on Markov chains. In an attempt to extend its applications to biopolymers, we introduce the propagation entropies of an electrostatic interaction using a protein road map. These negentropies are then used to classify Human Rhinoviruses as binding to the low-density lipoprotein receptor (100%) or intracellular adhesion molecule, with an accuracy of 88.89%. Overall accuracy varies between 91.6 and 100% in training and leave-group-out cross validation. In qualitative terms, the study opens the way to the application of topologic indices in biopolymer modelling.

Keywords: Markov chains; Human rhinovirus; Protein road maps

1. Introduction

Polymer chains and grid descriptors may be of great use when the aim of a research project is to derive quantitative polymer-structure property relationships. In relation to this, Arteca reported his mean crossing-over number as a descriptor for polymer chains accounting for the occurrence of entanglements caused by polymer chains interpenetrating each other [1,2]. Other molecular descriptors used to codify the molecular structure of polymers, including protein folding features, are the linking number, the Flory radius of gyration, I_3 index, and SDA (the sum of the cosines of dihedral angles) [3–8]. Describing the topology of polymers is also possible by using theoretical graphic and algebraic methods, which not only deal with linear polymers such as DNA, but also branched biopolymers, such as RNA. Randic's band average widths and series of sequence order coupling numbers encoding pseudo-amino acid composition describe linear biopolymers such as DNA and proteins [9–15].

In general, topologic indices have many other applications in the search for quantitative structure-activity relationships (QSAR) for small to medium-sized molecules, as well as for studying polymers [16,17]. Nevertheless, relationships between the structure and properties of polymers may also be sought using other techniques other than the topologic method. In particular, Markov chain theory has been found to have several applications in the field of biopolymers, particularly with proteins and nucleic acids [18–20].

It is also widely known that the surface of biopolymers may help in the recognition of its cellular receptors. As part of this process of connection, 'Road maps' are derived after a series of general steps. Firstly, a 2D projection is made of the 3D structure of the biopolymer. Monomers (amino acids) are then represented in a four-fold colour scale (blue-basic, red-acidic, yellow-hydrophobic, and green hydrophilic) [21,22]. Road maps have recently been applied by Vlasak et al. [23], in studying the receptor recognition of human Rhinoviruses (HRVs). In this study, a general rule was defined for predicting which of two possible receptors will be preferred by HRVs. HRVs are small, icosahedral

* Corresponding author. Address: Centro de Bioactivos Químicos, Diseño Computacional de Fármacos, Universidad Central de Las Villas, 5830 Santa Clara, Cuba. Tel.: +53-42-281473x131; fax: +53-42-281473x455.
E-mail address: humbertogd@vodafone.es (H. González Díaz), humbertogd@cbq.uclv.edu.cu (H. González Díaz).

0032-3861/$ - see front matter © 2004 Elsevier Ltd. All rights reserved.
doi:10.1016/j.polymer.2004.03.071

particles (~30 nm) composed of 60 copies of viral capsid proteins VP1, 2, 3, and 4 and a positive-strand (messenger sense) RNA, and may be classified into two groups: major group (viruses binding intracellular adhesion molecule 1, ICAM-1) and minor group (viruses binding low-density lipoprotein receptor, LDLR) [24–28]. Vlasak and co-workers' rule was derived after visually inspecting road maps. Unfortunately, the relationship between the structure of the biopolymers and the study property is not expected to ever be self-evident.

It was interesting for us to note in the present study that road maps have the appearance of a multicolored jigsaw. This means that they may be split into many pieces (amino acids), which are interconnected. This sentence can be automatically identified with the concept of coloured graphs, commonly treated in graph theory [29]. From here on, it is feasible to encode information about virus surfaces using theoretical graph invariants or topologic/topographic descriptors [30].

Taking these aspects into account, the present study is aimed at defining and using topologic descriptors to fit QSAR into the field of biopolymers. Here, we will extend our MARCH-INSIDE Markovian method [31–37], to predict rhinovirus receptors as an illustrative example of the application of topologic descriptors in modeling the properties of biopolymers.

2. Methods

2.1. Topologic Chapman–Kolmogorov decay of polar surface interactions

Consider an amino acid residue (aa_j) on the surface of the viral protein, interacting with an external factor (drug or viral receptor) at the initial time ($t_0 = 0$). Given one polarity driven by interaction, from here we will consider the problem of calculating the probability of this interaction affecting other amino acids (aa_k) at the viral envelope in future moments [31–37]. In accordance with the laws of thermodynamics, after the VP–receptor interaction, the aa_j of the VP forms a complex with the receptor in an excited state. Once this initial interaction takes place ($t_0 = 0$), it is expected that the more favorably the VP manages to fold to a stable complex, the higher the virus's preference for this receptor as compared to others [38,39]. Therefore, in the first stage the problem involves calculating the probability ($^0p_j(\varphi_j)$) of each aa_j having an initial interaction with the receptor at $t_0 = 0$. Secondly, attention could be focused on calculating the range of probabilities ($^kp_{ij}$) with which these interactions decay along the viral surface over time (t_k). This problem may be solved by making an analogy with former applications of the MARCH-INSIDE model [31–37]. This means that a number of simplifications have to be considered in order to develop suitable calculations. Firstly, in the case of calculations, we will only consider exposed aa_j as depicted in road maps. Secondly, the effect of the specific receptor is not considered in the probabilistic model but instead in the final statistical analysis. Eventually, an auxiliary proof charge (+1) is used to replace the viral receptor in all calculations. The general strategy is as follows:

(i) To develop a simple measurement (φ_j) for the polarity of aa_j.
(ii) To introduce the vector ($^0\Phi$) containing the initial absolute probabilities ($^Ap_0(\varphi_j)$) of polar interaction and the stochastic matrix ($^1\Pi$) representing the probabilities ($^1p_{ij}$) with which the initial interaction of aa_j (at time $t_0 = 0$) affects its closest neighbors at time $t_k = 1$, given a specific road map.
(iii) To define a stochastic process that generates the absolute probabilities ($^Ap_k(\varphi_j)$) with which the effect of the initial interaction ($t_0 = 0$) of the aa_i reaches aa_j at time $t_k = k$ (for details see: [31–37]).
(iv) To calculate the spectrum of time-dependent negentropies (negative entropies: $\Theta_k(S, \varphi_j)$) of $^1\Pi$ for specific regions (s) of the virus.
(v) Biophysical comments regarding the model will be explained in brief.
(vi) To use $\Theta_k(S, \varphi_j)$ as inputs in a Linear Discriminant Analysis (LDA) to seek a QSAR for receptors preferred by viruses [40].

2.1.1. The polarity of amino acids as characterized by φ_j

The electrostatic potential was selected at the molecular (amino acid) surface $\varphi_j = Q_j^*/R_j$ (aa_j-charge-radius ratio) due to its close relation with polarity [41]. In this case, R_j is the exposed amino acid radius (aa_j-radius). For the sake of simplicity, we consider only two radius levels: $R_j = 1$ for polar aa and $R_j = 2$ for non-polar levels.

Furthermore, Q_j^* is the corrected charge for the amino acid. The corrected charge is used to avoid non-positive values, bearing in mind that probabilities must not be negative, and the probability of electrostatic interaction of a neutral amino acid must not be 0, due to the possibility of induced charges [42]:

$$Q_j^* = Q_j + 3 : \ Q_j \equiv \begin{cases} Q_j = 1 & \text{for positively-charged-aa} \\ Q_j = 0 & \text{for non-charged-aa} \\ Q_j = -1 & \text{for negatively-charged-aa} \end{cases} \quad (1)$$

It is important to emphasize that despite its simplicity, the present polarity measurement (φ_j) may successfully account for the four classes of aa_j [21–23]. As the proof charge is considered with a fixed charge (+1) the electrostatic potential of interaction of an aa_j only depends on its charge and its radius.

(a) $\varphi_j = Q_j^*/R_j = 3/2 = 1.5$ for non-polar or hydrophobic aa_j (yellow = Y) <

(b) 2/1 = 2 for polar-negatively-charged or acid aa_j (Red = R) <
(c) 3/1 = 3 polar but non-charged aa_j (Green = G) <
(d) 4/1 = 4 positively charged aa_j (Blue = B).

2.1.2. Initial and secondary stages of the interaction between the virus surface and the receptor

The initial stage and secondary stage of the interaction between the virus and the auxiliary proof charge, representing the receptor, is described here using a probabilistic approach:

$$ {}^0\phi \equiv {}^0p_j(\varphi_j) = \frac{\varphi_j}{\sum_{k=1}^{n}\varphi_k} = \frac{Q_j^*/R_j}{\sum_{k=1}^{\delta+1}Q_j^*/R_k} \quad (2) $$

and

$$ \Pi_1 \equiv {}^1p_{ij} : {}^1p_{ij} = \frac{\varphi_j}{\sum_{k=1}^{\delta+1}\varphi_k} = \frac{Q_j^*/R_j}{\sum_{k=1}^{\delta+1}Q_j^*/R_k} \quad (3) $$

It is important to note that summation in (2) covers all the aa_j of the virus. On the contrary, summation in (3) only covers the aa_k of the virus that are adjacent to aa_i in the road map. Two different aa_j are only considered as adjacent to each other in the VRM if and only if the length of the contact frontier between them in the road map is >1. The interactions between aa_j with the contact frontier ≤1 are prohibited in the present model at times $t_0 = 0$ and $t_1 = 1$. In any case, these interactions may occur, indirectly, at a later time t_k ($k > 1$). This means that the propagation of the polar interaction follows the connectivity of the vertices in the graph derived from the road map. This fact determines the topological nature of the present model. It is very important to consider that the adjacency of two aa_j in the road maps does not imply that they are chemically bonded to each other, but instead that the exposed surfaces are neighbours. Furthermore, the present model is topological in terms of viral surface interactions, but contains real 3D information about the viral structure. In fact, road maps are derived from X-ray, NMR or 3D-computational models [36,37,43].

2.1.3. Markov chain processes for polar virus–receptor interaction decay

The most important aspect involved in the present study is the consideration that once viral aa_j-receptor interaction takes place; its propagation obeys the Chapman–Kolmogorov equations [44]. In mathematical terms, the vectors ${}^k\Phi$ with elements equal to the absolute time-dependent probabilities (^kp_j), with which each aa_j influences the folding kinetics of the complex virus–receptors, depends on the natural powers of ${}^1\Pi$:

$$ {}^k\phi \equiv {}^kp_j(\varphi_j) : {}^k\phi = {}^0\phi \times (\Pi_1)^k \quad (4) $$

This hypothesis makes it possible to calculate all time-dependent probabilities, only having the initial probability vector ($^0\Phi$) and the stochastic matrix ($^1\Pi$). This means that the process of relaxation of the virus envelope after the initial interaction with the receptor may be represented as a Markov chain. As a result, the initial time probabilities (at $t_0 = 0$) and the second-stage probabilities at $t_1 = 1$ (represented by the matrix $^1\Pi$) govern the folding kinetics of the virus–receptor complex at subsequent times $t_k = 2$, 3, 4, 5… until reaching the stationary state [31–37].

2.1.4. Local and total negentropies of time dependent polar–surface interactions as molecular descriptors of the viral structure

Molecular negentropies [45,30] have proved to be an excellent source for the definition of novel molecular descriptors. With regard to this, our research group has investigated the application of novel types of molecular negentropies $\Theta_k(s, w_j)$ in order to study the structure-property relationships of proteins and nucleic acids [33,37,44,46,47].

$$ \Theta_k(s, w_j) = -k_B \sum_{n=0}^{s} {}^Ap_k(w_j)\log {}^Ap_k(w_j) \quad (5) $$

Where w_j is a weight characterizing the phenomena under study [31–37], e.g.:

$w_j = \chi_j$, Pauling electronegativity, or any other function
$w_j = f(n_{ij})$, being n_{ij} the number of shared electrons between two atoms.
$w_j = \nu_j$ represents the frequency (energy) of an elastic vibration to describe its probability of propagation through a nucleic acid backbone (DNA or RNA).
$w_j = ECI_j$ the electronic charge index of the amino acid to encode the protein structure.
$w_j = \varphi_j$ in this paper.

Considering the atoms, nucleotides, and amino acids of the VP as independent sources of entropy [33], the sum of $\Theta_k(s, w_j)$ for a set of elements (s) at the same t_k and with the same w_j constitute the entropy related to the phenomenon characterised by w_j, at this t_k for this collection of elements.

In specific terms, if the collection s contains the sum of the vertices (aa_j in the present work) then $s = T$ and $\Theta_k(T, \varphi_j)$ becomes a total descriptor. Conversely, if the collection s contains only one specific vertex (e.g. the Lys in the HI loop of the HRVs) then $s =$ Lys and $\Theta_k(L, \varphi_j)$ becomes a local descriptor. The HI loop is a specific feature present in HRVs, whose information content has been proposed as being of utmost importance in receptor recognition [48]. Amino acid class grouped by indices are another example of the local indices obtained, for example: $s = B$ (basic aa_j). The calculation of the virus surface descriptors described above has been implemented in our

user-friendly software application, MARCH-INSIDE 2.0 [49].

2.1.5. Biophysical comments on the model

One of the main problems in developing physical and biophysical theories is selecting the invariant scales for the units. Labeling, chemical, translational, rotational, and conformational invariance are the most widely used scales for molecular descriptors [50]. With regard to R_j, the elements of the stochastic matrix and therefore the viral descriptors ($\Theta_k(S, \varphi_j)$) derived are invariant under linear changes of scale $'R_j = \alpha R_j$. This means that the probabilities calculated with the former scale ($^1p_{ij}$) are identical to the pattern of probabilities calculated with each alternative scale ($'^1p_{ij}$) as follows:

$$^1p_{ij} = \frac{\varphi_j}{\sum_{k=1}^{\delta+1}\varphi_k} = \frac{Q_j^*/'R_j}{\sum_{k=1}^{\delta+1}Q_j^*/'R_k} = \frac{Q_j^*/\alpha' R_j}{\sum_{k=1}^{\delta+1}Q_j^*/\alpha R_j}$$

$$= \frac{\frac{1}{\alpha}Q_j^*/R_j}{\frac{1}{\alpha}\sum_{k=1}^{\delta+1}Q_j^*/R_k} = \frac{Q_j^*/R_j}{\sum_{k=1}^{\delta+1}Q_j^*/R_k} = {}^1p_{ij} \quad (6)$$

Where $'R_j$ is the new radio scale and α represents the conversion factor from one scale to other (R_j). Accordingly, the stochastic matrix is also invariant with regard to the selection of the Q_j^* scale. This may be easily demonstrated by substituting Q_j^* with another scale related to it linearly by the conversion factor β ($'Q_j^* = \beta Q_j^*$).

In this case, there are some physical reasons for selecting a topological Chapman–Kolmogorov behaviour. Firstly, virus–receptor interaction is a docking-like molecular phenomenon and as this must obey the laws of quantum or at least molecular mechanics [38,39]. This means that the stochastic process selected must therefore obey Heisenberg's uncertainty, and the principle of indistinguishability of identical particles (electrons) [51]. The selection of Markov chains determines that the present model does not depend on the labeling of electrons at previous configurations (less memory) [52,53]. On the other hand, this imposes the fact that it is not possible to precisely determine the propagation of the interaction, only its probabilities [31–37].

Another interesting property of this procedure refers to the algebraic properties traditionally established in mathematical biology for the matrix representation of biopolymers. If we consider M as the matrix representing the 2D structure of the biopolymer and m as a vector that codifies information about its monomers (nucleotides or aa), they are both traditionally compelled to obey the following self-consistent relationship: $m \times M = m$ [10]. The present methodology does not only uphold this condition, but also lends it profound biophysical significance. A classical result of the theory of Markov chains is the limiting behaviour of $^Ap_k(\varphi_j)$ when k tends to be infinite (∞) [20,44,52,53]. In mathematical terms, a matrix such as $^1\Pi$ and the vector $^\infty\Phi$ are interconnected by the following relationship: $^\infty\Phi \times {}^1\Pi = {}^\infty\Phi$. This implies that $^\infty\Phi$ is an Eigenvector $^1\Pi$ with elements ($^Ap_\infty(\varphi_j)$). Therefore, the initial aa_j-receptor interaction affects the other aa_k at the virus surface, with time changing absolute probabilities ($^Ap_k(\varphi_j)$), which reach limiting or stationary values $^Ap_\infty(\varphi_j)$ at time t_∞. In biophysical terms, this means that the folding change originated at the virus surface by the initial polar interaction is not infinite. Conversely, this folding process reaches the stationary state in a quasi-static manner. Actually, the vectors $^k\Phi$ are highly collinear or otherwise non-orthogonal ($^{k-1}\Phi \times {}^k\Phi' \neq 0$) for quite short periods of time: $t_k = k \sim 15$. In practical terms, this means that the numerical calculations achieve results that are very similar to the stationary results at $t_\infty = k \ll \infty$. In the previous analysis, the symbol $^k\Phi'$ refers to the transposition of the vectors $^k\Phi$.

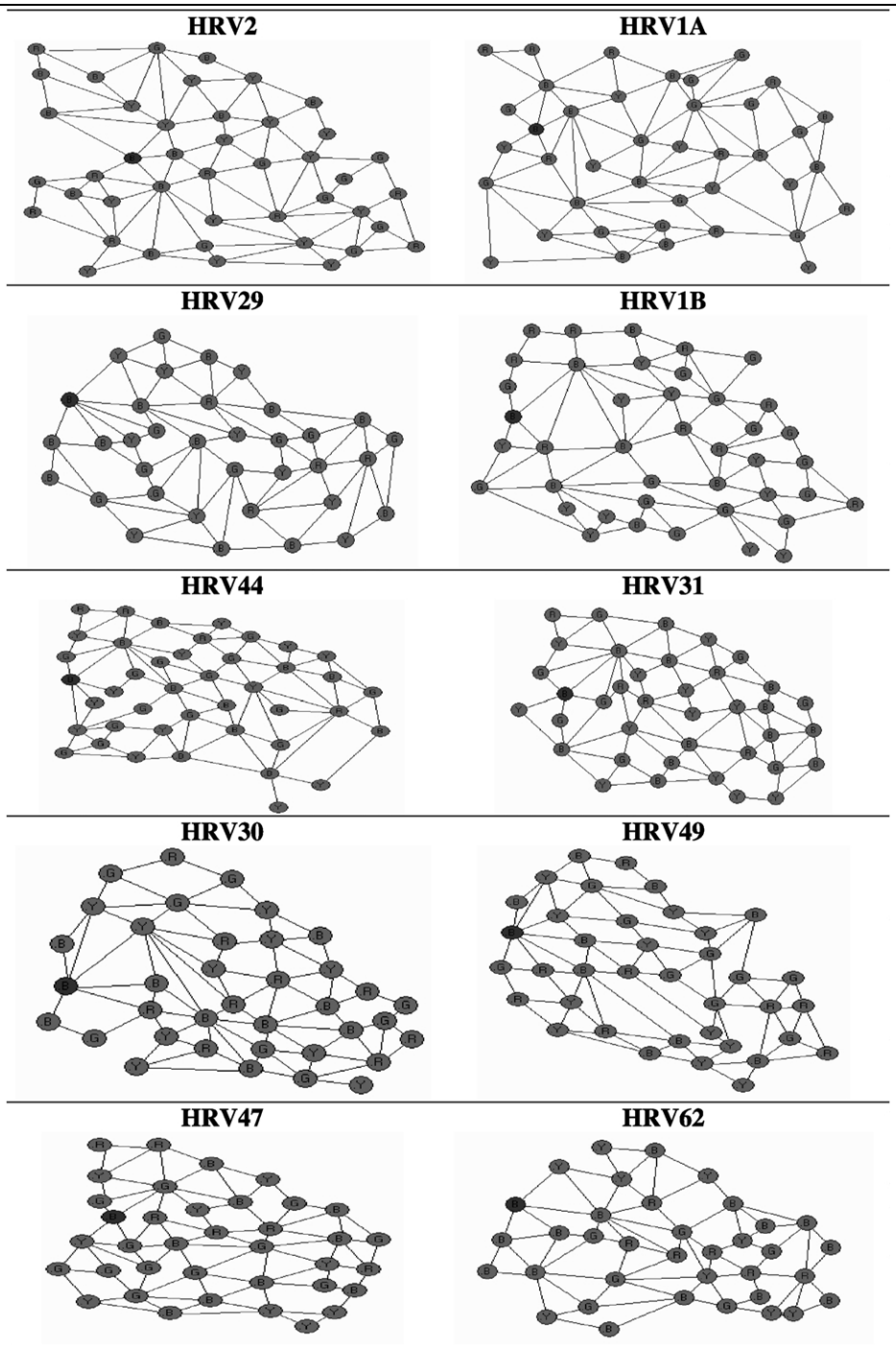
2.1.6. Stochastic negentropies and viral property relationships

In the present study, LDA is proposed for use in seeking a linear discriminatory function to predict which of two possible receptors are preferred by the virus [54,55]. This relationship may be represented using the following general expression:

$$Vr = b_0 + b_{1k}\Theta_k(L, \varphi_j) + b_{2Bk}\Theta_k(B, \varphi_j) + b_{3k}\Theta_k(T, \varphi_j) \quad (7)$$

Where, Vr (viral receptor), the output of the model, is a dummy variable (Vr = 1 if the virus receptor is LDLR and −1 if it is ICAM-1). In the equation, b_0, b_{1k}, b_{2Bk}, and b_{3k} are the stepwise LDA coefficients as estimated using the STATISTICA 6.0 software package [56]. These coefficients may account for information about the direct influence of the negentropies calculated considering the proof-charge vs. virus interaction with regard to the real system constituted by the pair-wise interaction between the virus and its receptor. In this equation, $\Theta_k(B, \varphi_j)$, $\Theta_k(L, \varphi_j)$, and $\Theta_k(T, \varphi_j)$ act as viral descriptors and constitute the inputs of the model. The training quality of this model was assessed by direct inspection of different statistics such as good classification percentages (% LDLR, % ICAM, % Total), Wilks's statistics (U), the Fisher ratio (F) and the probability of error (p-level (p)). The parameters % LDLR and % ICAM are good classification percentages for viruses binding to one of the two possible viral receptors, LDLR and ICAM. Also, % Total is the total good classification percentage. The quality of the model was considered as acceptable if all of these percentages were >85%. Statistical significance was measured by selecting models whose values for U and F imply that $p < 0.05$ [31–37]. Furthermore, the model was validated by carrying out resubstitution experiments. A total of four leave-group-out runs were performed. In doing so, 1 out of every 4 compounds was extracted at random. The model with the

Table 1
Minor group HRVs 4-folded-colored-graphs



smallest variation for all of the parameters with regard to the training series and cross-validation was selected [57].

3. Results and discussion

The same HRVs series recently used by Ref. [23] was considered here as a training series. A total of 19 HRVs were studied: 10 belonging to minor group and the other 9 to the major group. Initially, the surface structure of all the HRVs was entered into the MARCH-INSIDE 2.0 software as 4-folded coloured graphs. These graphs were derived from the Road maps as detailed in Section 2.1.2 of this article. The respective graphs for each one of the minor group HRVs are shown in Table 1.

Also, the graphs for major group HRVs are shown in

Table 2
Major group HRVs 4-folded-colored-graphs

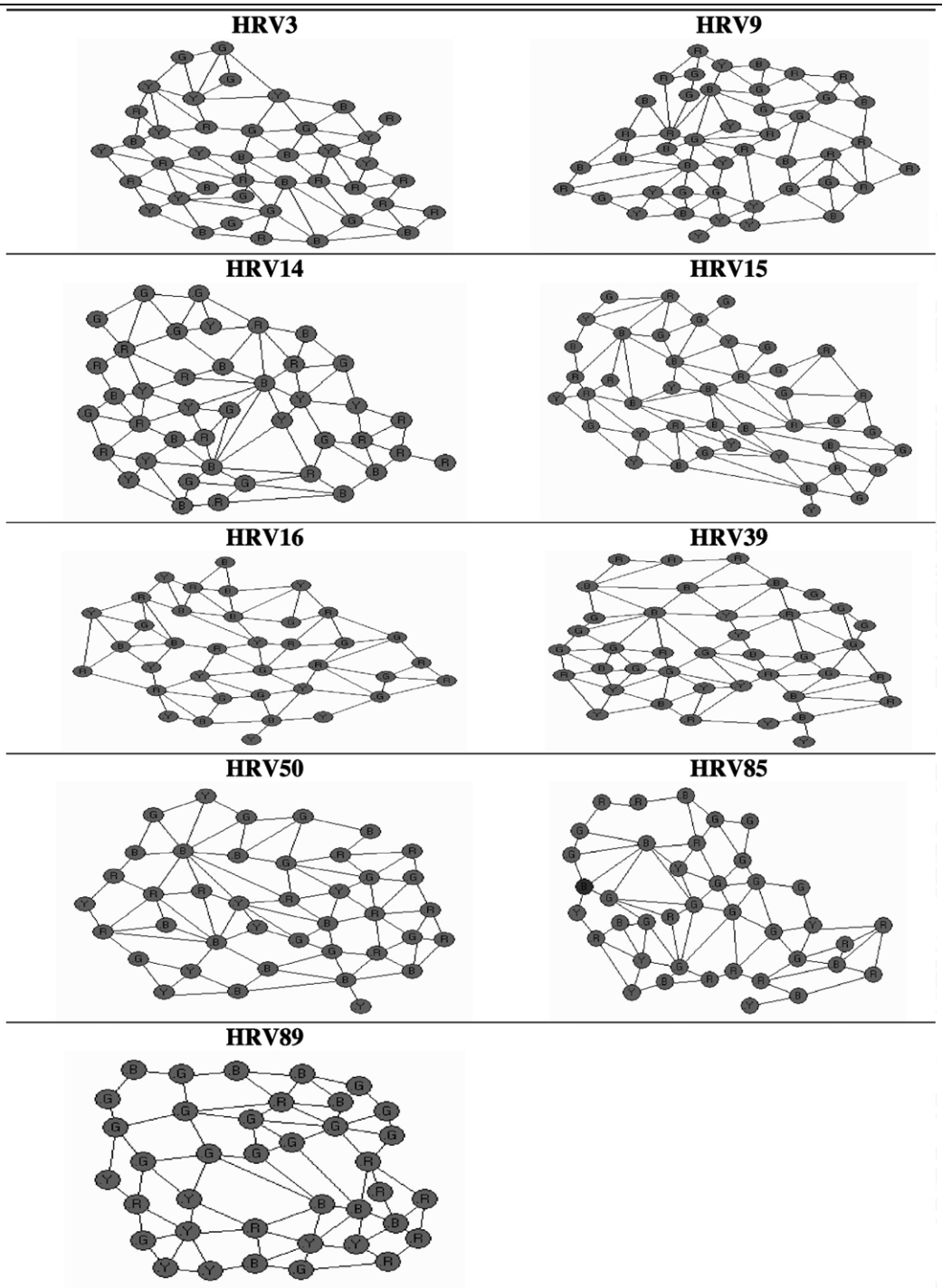


Table 2. Secondly, calculations were made of the six first negentropies ($t_k = 0, 1, 2 \ldots 5$) for the three different kinds of viral surface regions $\Theta_k(B, \varphi_j)$, $\Theta_k(L, \varphi_j)$, and $\Theta_k(T, \varphi_j)$. The significance of $\Theta_k(B, \varphi_j)$, $\Theta_k(L, \varphi_j)$, and $\Theta_k(T, \varphi_j)$ stay the same; they are the Markovian propagation negentropies with the time of the polar interaction between the amino acids and the proof charge. Three sets of amino acids were considered: L = Lysine in the HI loop, B = Basic amino acids, and T = all the amino acids of the VP.

An LDA was then performed in order to detect the model. The best model found included only three variables $\Theta_0(B, \varphi_j)$, $\Theta_0(L, \varphi_j)$, and $\Theta_5(T, \varphi_j)$. Therefore, the model has an acceptable balance (ρ) between the number of cases (N) and the number of adjustable parameters (N_{ap}). For the

Table 3
Summary of statistical analysis for training and Jack-knife cross validation

	ALL HRV	CV1	CV2	CV3	CV4
n	19	13	12	13	13
u	0.17	0.16	0.24	0.07	0.16
f	24.8	15.6	8.24	419.5	15.5
p	<0.00	<0.00	<0.00	<0.00	<0.00
%LDLR	100	100	83.33	100	100
%ICAM	88.89	100	100	100	100
%Total	94.7	100	91.66	100	100

present specific case $N_{ap} = 8$ due to the fact that we are dealing with a two-group LDA. This means that the number of discriminant functions is $N_{cf} = 1$, and taking this into account, each function has 3 variables ($N_v = 3$) and 1 intercept to be estimated: $\rho = N/N_{ap} = N/N^*_{cf}(N_v + 1) = 19/1 \times (3 + 1) = 3.16$, which is an acceptable value [54]. The equation is presented below, and was selected by considering all the aspects explained in section 2.2:

$$\mathrm{Vr} = -22.526 + 289.943 \times \Theta_0(\mathrm{L}, \varphi_j) + 24.074 \times \Theta_0(\mathrm{B}, \varphi_j) + 0.841 \times \Theta_5(\mathrm{T}, \varphi_j) \quad (8)$$

Details of the training and cross-validation properties of the model are presented in Table 3. As may be seen, the values of F and U statistics were sufficient to ensure the statistical significance of the model for training, as well as for cross-validation. In all cases the p-level was <0.00, and furthermore the model has shown to have very good predictability in both training series, ranging from 88.89% to 100%. The 3 variables may be clearly identified with 3 factors governing virus docking to the molecular receptor. The first, and strongest factor, $\Theta_0(\mathrm{L}, \varphi_j)$ codifies information about the number of Lys residues in the HI loop. The highly positive influence of this factor coincides with very well documented results in the literature [58]. The second most important variable according to the statistical results is $\Theta_0(\mathrm{B}, \varphi_j)$, which is directly related to the number of basic amino acids of the VP exposed at the surface of the HRVs. This result confirms the previous observations of Vlasak et al. [23] with regard to the importance of the pattern of basic amino acids on the viral surface. Finally, the model detected a less important but still significant factor $\Theta_5(\mathrm{T}, \varphi_j)$. This variable represents the entropy of the process of propagation of the dipolar interaction on the entire viral surface until $t_k = 5$. This quantity therefore codifies middle-to-long-term folding of the entire viral surface after the initial interaction. For this reason, this variable codifies information about the whole structure of the VP as being the most susceptible to phylogenetic changes with respect to the other two variables. The other two are local variables. In this sense, is natural to consider $\Theta_5(\mathrm{T}, \varphi_j)$ as a variable that determines the specific probability of binding once $\Theta_0(\mathrm{B}, \varphi_j)$, and $\Theta_0(\mathrm{L}, \varphi_j)$ as discerning the virus receptor. Nevertheless, it is important to emphasize that the local variables alone do not seem to be capable of discriminating the HRVs receptor in all cases, and require the complementary information from the total receptor.

The interpretation offered in this article agrees with the statistical behavior of the model's robustness and the subsequent probability of receptor interaction predicted for each HRV. In Table 3, the lowest accuracy appeared in the CV2 experiment (83.33%). Nevertheless, these are good enough

Table 4
Results of the LDA classification in training and leave-group-out cross validation

HRV	Observed virus receptor	Prob	Probcv1	Probcv2	Probcv3	Probcv4
2	LDLR	0.999	0.998	0.990[a]	1.000	0.998
1A	LDLR	1.000	0.997	0.994	0.999[a]	0.997
1B	LDLR	0.985	0.101[b]	0.821	1.000	0.101[a]
29	LDLR	1.000	1.000	0.999[a]	1.000	1.000
30	LDLR	0.998	0.975	0.982	0.999[a]	0.975
31	LDLR	1.000	0.999[a]	1.000	1.000	0.999[a]
44	LDLR	1.000	1.000	0.999[a]	1.000	1.000
47	LDLR	1.000	1.000	1.000	0.999[a]	1.000
49	LDLR	0.999	0.994[a]	0.995	1.000	0.994[a]
62	LDLR	1.000	1.000	0.999[a]	1.000	1.000
3	ICAM	1.000	1.000	0.999[a]	1.000	1.000
9	ICAM	1.000	1.000	0.999	0.999[a]	1.000
14	ICAM	1.000	0.999[a]	0.999	1.000	0.999[a]
15	ICAM	1.000	1.000	0.998[a]	1.000	1.000
16	ICAM	1.000	1.000	0.999	0.999[a]	1.000
39	ICAM	1.000	0.993[a]	0.998	1.000	0.993[a]
50	ICAM	1.000	1.000	0.991[a]	1.000	1.000
85	ICAM	0.005[b]	0.617	0.065[b]	0.000[a,b]	0.617
89	ICAM	1.000	0.992[a]	0.998	1.000	0.992[a]

[a] Left-out-compound in Jackknife cross-validation.
[b] Misclassified compound.

percentages for one interesting case to be discussed. Virus serotype 85, which is the cause of a lack of accuracy in CV2 and is also misclassified in training, and CV3 (compare Tables 3 and 4), which seems to be an outlier.

However, it is important to note that this is not merely a statistical outlier. There are certainly biological motives for considering this serotype as special, having sequential and phylogenetic characteristics that are common to the major group, but 3D features and receptor affinities like those of a minor group member [59].

4. Conclusions

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

1. Topologic descriptors may be used to predict the properties of biopolymers from the properties of their 3D-surface.
2. The graph derived from the biopolymer road map may act as the source for these descriptors.
3. In particular, the 3D generalization of MARCH-INSIDE for predicting Human Rhinovirus receptors illustrates how the method may be used in similar situations in polymer sciences.

Acknowledgements

Gonzalez DH would like to express his sincere gratitude to Dr Jose Luis Garcia and The Cuban Ministry of Higher Education for its partial financial support. González DH and Uriarte E would also like to thank the Xunta of Galicia for grant PR405A2001/65-0, which was used to purchase the Statistica 6.0 software. The same authors also offer their thanks to the Spanish Ministry of Science and Technology (SAF2003-02222), for its partial financial support.

References

[1] Arteca GA. J Chem Inf Comput Sci 1999;39:550.
[2] Arteca GA, Mezey PG. J Mol Graphics 1990;8:66.
[3] White JH. Am J Math 1969;91:693.
[4] Fuller FB. Proc Nat Acad Sci USA 1971;68:815.
[5] Flory PJ. Principles of polymer chemistry. Itaha: Cornell University Press; 1953.
[6] Fresht A. Structure and mechanism in protein science. New York: W.H. Freeman; 1999.
[7] Estrada E. Bioinformatics 2002;18:1.
[8] Estrada E. Chem Phys Lett 2000;319:713.
[9] Roy A, Raychaudhury C, Nandy A. J Biosci 1998;23:55.
[10] Casanovas J, Miro-Julia J, Rosselló F. J Math Biol 2003;47:1. and references cited therein.
[11] Leong PM, Mogenthaler S. Comput Appl Biosci 1995;12:503.
[12] Randić M, Vračko M, Nandy A, Basak SC. J Chem Inf Comput Sci 2000;40:1235.
[13] Randić M, Balaban AT. J Chem Inf Comput Sci 2003;43:532.
[14] Hua S, Sun Z. Bioinformatics 2001;17:721–8.
[15] Cai Y-D, Lina SL. BBA 2003;1648:127.
[16] Randić M. J Chem Inf Comput Sci 1997;37:672.
[17] Hall LH, Mohney B, Kier LB. J Chem Inf Comput Sci 1991;31:76.
[18] Chou KC. Biopolymers 1997;42:837.
[19] Vorodovsky M, MacIninch JD, Koonin EV, Rudd KE, Médigue C, Danchin A. Nucleic Acids Res 1995;23:3554.
[20] Yuan Z. FEBS Lett 1999;451:23.
[21] Chapman MS. Protein Sci 1993;2:459.
[22] Chapman MS, Rossman MG. Virology 1993;195:745.
[23] Vlasak M, Blomqvist S, Hovi T, Hewat E, Blaas D. J Virol 2003;77:6923.
[24] Andries KB, Denwindt B, Snoeks J, Wouters L, Moereels H, Lewi PJ, Janssen PAJ. J Virol 1990;64:1117.
[25] Register RB, Uncapher CR, Naylor AM, Lineberg DW, Colonno RJ. J Virol 1991;65:6589.
[26] Howell BW, Herz J. Curr Opin Neurobiol 2001;11:74.
[27] Herz J. Nat Struct Biol 2001;8:476.
[28] Ruecker RR. Picornaviridae: the viruses and their replication. In: Fields BN, Knipe DM, Howley PM,, editors, 3rd ed. Fields virology, vol. 1. Philadelphia, PA: Lippincott-Raven Publishers; 1996. p. 609–54.
[29] Trinajstic N. Chemical graph theory. Boca Raton, Florida: CRC Press; 1992. p. 322.
[30] Todeschini R, Consonni V. Handbook of molecular descriptors. Wemheim, Germany: Wiley-VCH; 2000. p. 667.
[31] González DH, Olazábal E, Castañedo N, Hernádez SI, Morales A, Serrano HS, González J, Ramos de Armas R. J Mol Mod 2002;8:237.
[32] González DH, Hernández SI, Uriarte E, Santana L. Comput Biol Chem 2003;27:217.
[33] González DH, Marrero Y, Hernández I, Bastida I, Tenorio I, Nasco O, Uriarte E, Castañedo N, Cabrera MA, Aguila E, Marrero O, Morales A, Pérez M. Chem Res Toxicol 2003;16:1318.
[34] González 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. J Mol Mod 2003;9:395.
[35] González DH, Ramos de AR, Uriarte E. Online J Bioinf 2002;1:83.
[36] González DH, Ramos de AR, Molina R. Bull Math Biol 2003;65:991.
[37] González DH, Ramos de AR, Molina R. Bioinformatics 2003;19:2079.
[38] Wang R, Lu Y, Wang S. J Med Chem 2003;46:2287.
[39] Wang R, Liu L, Lai L, Tang Y. J Mol Mod 1998;4:379.
[40] González MP, González DH, Molina R, Cabrera MA, de Armas RR. J Chem Inf Comput Sci 2003;43:1192.
[41] Bonaccorsi R, Scrocco E, Tomasi J. J Chem Phys 1970;52:5270.
[42] Andrews PR. Drug-receptor interactions. In: Kubinyi H, editor. 3D QSAR in drug design. Theory, methods and applications. Leiden, The Netherlands: ESCOM; 1993. p. 13–40.
[43] Guex N, Peitsch MC. Electrophoresis 1997;18:2714.
[44] Freund JA, Poschel T. Stochastic processes in physics, chemistry and biology. Lecture notes in physics. Berlin: Springer; 2000. p. 1–49.
[45] Kier LB. J Pharm Sci 1980;70:583.
[46] Krogh A, Brown M, Mian IS, Sjeander K, Haussler D. J Mol Biol 1994;235:1501.
[47] Shannon C. Bell Sys Tech J 1948;27:379.
[48] Duechler MS, Ketter S, Skern T, Kuechler E, Blaas D. J Gen Virol 1993;74:2287.
[49] González DH, Molina R, Hernández I. MARCH-INSIDE version 2.0, (MARkovian CHemicals 'In SIlico', DEsign). 2003. Chemicals Bioactives Center, Central University of 'Las Villas', Cuba. This is a preliminary experimental version, details upon request to the corresponding author at: humbertogd@cbq.uclv.edu.cu or humbertogd@navegalia.com.
[50] Charton M. The upsilon steric parameter X: definition and determination. In: Charton M, Motoc I, editors. Steric effects in drug design. Topics in current chemistry, vol. 114. Berlin, Germany: Springer-Verlag; 1983. p. 57–91.

[51] Landau LD, Lifshitz EM. Mecánica quántica no-relativista. Curso de física teórica, vol. 3. Barcelona: Reverté; 1963. p. 1–49.
[52] Bharucha-Reid AT. Elements of theory of Markov process on the plication. McGraw-Hill series in probability and statistics. New York: McGraw-Hill Book Company; 1960. p. 167–434.
[53] Van Kampen NG. Stochastic process in physics and chemistry. New York: North-Holland; 1981.
[54] García-Domenech R, de Julian-Ortiz JV. J Chem Inf Comput Sci 1998;38:445. and references cited therein.
[55] Gálvez J, García-Domenech R, De Julian-Ortiz V, Soler R. J Chem Inf Comput Sci 1994;34:1198.
[56] Statsoft Inc., 2002. STATISTICA for windows, Version 6.0.
[57] Efron B. J Am Stat Asoc 1983;78:316.
[58] Colonno RJ, Condra JH, Mizutani S, Callahan PL, Davies ME, Murcko MA. Proc Nat Acad Sci USA 1988;85:5449.
[59] Uncapher CR, Dewit CM, Colonno RJ. Virology 1991;180:814.