Available online at www.sciencedirect.com
SCIENCE DIRECT®
Bioorganic & Medicinal Chemistry Letters 14 (2004) 4691–4695
Bioorganic & Medicinal Chemistry Letters

Markov entropy backbone electrostatic descriptors for predicting proteins biological activity

Humberto González-Díaz,[a,b,*] Reinaldo Molina[a,c] 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 15782, Spain*
[c]*Universität Rostock, FB Chemie, Albert-Einstein-Str. 3a, D 18059 Rostock, Germany*

Received 8 April 2004; revised 25 June 2004; accepted 28 June 2004
Available online 24 July 2004

Abstract—The spherical truncation of electrostatic interactions between aminoacids makes it possible to break down long-range spatial electrostatic interactions, resulting in short-range interactions. As a result, a Markov Chain model may be used to calculate the probabilities with which the effect of a given interaction reaches aminoacids at different distances within the backbone. The entropies of a Markov Chain model of this type may then be used to codify information about the spatial distribution of charges in the protein used in this study exploring the structure–activity relationship. In this paper, a linear discriminant analysis is reported, which correctly classified 92.3% of 26 proteins under investigation and leave-one-out cross validation, purely for illustrative purposes. Classification was carried out for three possible activities: lysozymes, dihydrofolate reductases, and alcohol dehydrogenases. The discriminant analysis equations were contracted into two canonical roots. These simple canonical roots have high regression coefficients ($R_{c1} = 0.903$ and $R_{c2} = 0.70$). Root1 explains the biological activity of alcohol dehydrogenases while Root2 discriminates between lysozymes and dihydrofolate reductases. It was possible to profile the effect of core, middle, and surface aminoacids on biological activity. In contrast, a model considering classic physicochemical parameters such as: polarizability, refractivity, and partition coefficient classify correctly only the 80.8% of the proteins.

The search for molecular descriptors to discover quantitative structure–activity relationships (QSAR) for proteins may to some extent be considered as an emerging field, when compared with the interconnection that exists between macromolecular, bioorganic, medicinal, and computational chemistry. In this sense, some useful molecular descriptors have appeared for proteins and other polymers such as: mean over crossing number, the linking number, the Flory radius of gyration, the I3 index, SDA (sum of cosines of dihedral angles) and sequence order coupling numbers.[1–7] Some authors have also used various α-helix-propensity descriptors such as the Einserberg, Garnier, and Chou–Fasman scales; others have measured hydrophobicity (Kyle–Dolittle hydrophobicity and mean hydrophobicity moment) and surface (such as the Emini Surface Index).[8] All these new molecular descriptors have appeared in addition and as a complement to those traditionally at the disposition of researchers such as physicochemical parameters related to polarizability, refractivity, and partition coefficients. However, the number of molecular descriptors reported for proteins is quite low if compared with the large number of molecular descriptors developed for small-to-medium sized molecules. In the field of proteins, considerably less research has been carried out into the molecular descriptors based on entropic effects related to electrostatic and hydrophobic interactions.[9]

Our research group has developed simple stochastic molecular descriptors based on the Markov Chain (MC) theory. They describe changes in the electron distribution and the propagation of vibrations throughout the molecular backbone. Applications included the prediction of the flukicidal and anticancer activity of novel drugs.[10,11] Promising results have also been obtained in modeling the interaction between drugs and the HIV packaging region RNA in the field of bioinformatics.[12] However, the prediction of protein properties has only been addressed quite recently using this approach.[13] Although the codification of chirality and other 3D structural features is an advantage of this method,[14] no reports

Keywords: Protein electrostatics; 3D-QSAR; Markov chain; Electrostatic field.
* Corresponding author. Tel.: +534-228-1473; fax: +534-228-1130; e-mail addresses: humbertogd@vodafone.es; humbertogd@uclv.edu.cu

0960-894X/$ - see front matter © 2004 Elsevier Ltd. All rights reserved.
doi:10.1016/j.bmcl.2004.06.100

exist of any previous attempt to explicitly account for 3D structure. Offering a feasible interpretation in physical terms using the entropy concept is one of the most interesting advantages of this approach.[15–19] This paper, therefore, explores extending the previously mentioned stochastic molecular descriptors for 3D-QSAR in proteins, considering entropic effects related to electrostatic fields.

Consider a representation for a protein described as a static model, which considers a spatial distribution of pseudo aminoacids with 3D coordinates (x_i, y_i, z_i) coinciding with those for the respective Cα. In this case, every pair of aminoacids (i,j) present 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 as reported by Collantes and Dunn.[20] As a result, it is then easy to deal with the problem of the propagation of the effect of all aminoacid–aminoacid pairwise electrostatic interactions within the protein backbone. 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 ${}^1\Pi(x,y,z,q)$.[10,11,14,21] 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 (${}^1pij$) with which the aminoacid i presents a 'truncated' electrostatic interaction of energy E_{ij}, with the aminoacid j placed at a distance d_{ij}:[18]

$$ {}^1p_{ij} = \frac{E_{ij}}{\sum_{k=1}^{n_j} E_{ik}} = \frac{\delta_{ij} \cdot q_j / d_{ij}^2}{\sum_{k=1}^{n_j} \delta_{ik} \cdot q_j / d_{ij}^2} \qquad (1) $$

In our model, the 'truncated' electrostatic interactions are those, which occur between aminoacids at a cut-off distance (d_{in}) shorter than the half of the sum of their van der Waals radius. Here, a shifting function (δ_{ij}) was used to indicate the presence $(\delta_{ij} = 1)$ or absence $(\delta_{ij} = 0)$ of the 'truncated' electrostatic interaction between aminoacids i and j. This procedure forms part of what it is known as truncation methods, and is referred to as the spherical truncation method. These methods neglect all Coulomb or van der Waals interactions where the distance between the two atoms is greater than a certain cut-off distance.[22]

In Eq. (1) the sum was made of all the n_j aminoacids that had spherical truncated interactions with the aminoacid i. In other words, in this study the electrostatic field was transformed from a continuous field to a discrete field, making a direct MC matrix codification possible. The main approximation here is undoubtedly to consider that a spherical truncated electrostatic interaction may propagate throughout space to other aminoacids in the protein as a MC. The present approach neglects long-range Coulomb interactions (dotted arrow) in the stochastic matrix but conversely to classic truncation methods allows for them in a step-by-step fashion (solid arrows), as shown in Figure 1.

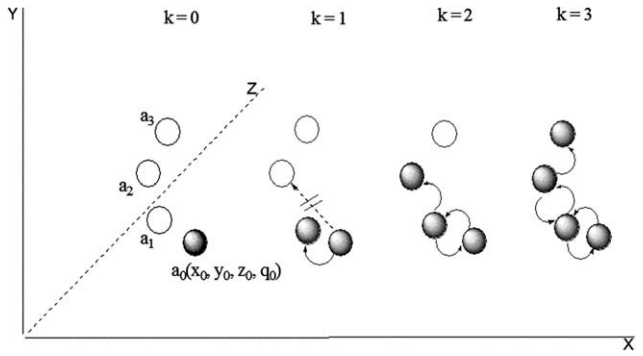

Figure 1. Spatial propagation of Markov truncated electrostatic interaction between one aminoacid and the others at the protein 3D backbone.

The spatial dependence of the model becomes clear on inspection of Figure 1. Due to truncation restrictions, the aminoacid a_0 (with charge q_0 and coordinates x_0, y_0, z_0) is only able to interact with the aminoacid a_1 by means of direct interaction. The effect of this interaction can only affect aminoacid a_2 in a subsequent propagation of the interaction a_1-a_2 and so on. It is clearer here identifying the parameter of the MC as 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 are given at corresponding discrete time intervals $(\Delta t_k = t_{k+1} - t_k = k)$ like in almost MC applications.

We will consider that the absolute probabilities (${}^{\text{A}}p_k(j)$) with which these long-range interactions rise to a specific aminoacid j after k steps are the elements of the vectors ${}^k\Phi(x,y,z,q)$ derived in a Markovian manner using the so-called Chapman–Kolmogorov equations:[11–13,21]

$$ \begin{aligned} {}^k\phi(x,y,z,q) &= {}^0\phi(x,y,z,q) \times {}^k\Pi(x,y,z,q) \\ &= {}^0\phi(x,y,z,q) \times \left[{}^1\Pi(x,y,z,q)\right]^k \end{aligned} \qquad (2) $$

where ${}^0\phi(x,y,z,q)$ is the vector in which elements (${}^{\text{A}}p_0(j)$) are the initial absolute probabilities with which any aminoacid j participates in an electrostatic.[18]

After this simple approximation, calculating the spectrum of entropies with which the electrostatic interaction propagates throughout the protein backbone is relatively straightforward:[16–19,21]

$$ \begin{aligned} \Theta_k(\text{orbit, pol}) &= -k_{\text{B}} \sum_{j=0}^{n} {}^{\text{A}}p_k(j) \log {}^{\text{A}}p(j) \\ &= -k_{\text{B}} \sum_{S} {}^{\text{A}}p_k(S) \log {}^{\text{A}}p(S) \\ \Theta_k(\text{orbit, pol}) &= -k_{\text{B}} \sum_{\text{orbit}} \sum_{\text{pol}} {}^{\text{A}}p_k(\text{orbit, pol}) \log {}^{\text{A}}p(\text{orbit, pol}) \end{aligned} \qquad (3) $$

where the collection of all the $j = 1$ to n aminoacids in the protein can be regrouped into different sub-sets (S)

of aminoacids. These S-sub-sets can be based on different combinations with respect to the parameter polarity of the aminoacids (pol = 'polar', 'non-polar', 'all aminoacids') and/or the orbit they occupy with respect to the center of mass of the protein. Here, five values were used by default for the parameter orbit = 0, 1, 2, 3, 4′, considering aminoacids with a ratio $r = d(j)/d_{max}(j)*100$ ranging between the following limits: $0 \leqslant \text{orbit} = 0 \leqslant 25 \leqslant \text{orbit} = 1 \leqslant 50 \leqslant \text{orbit} = 2 \leqslant 75 \leqslant \text{orbit} = 3 \leqslant 100\%$ or all of the aminoacids together, orbit = 4. This partition makes it possible for us to calculate values such as the entropy Θ_5 (orbit3, non-polar) of the electrostatic interactions involving non-polar aminoacids placed at a topologic distance $k = 5$ and near the surface of the protein. That is to say, aminoacids at a topological distance (number of short-range interactions) equal to 5. As the orbits are related to the position of the aminoacid with respect to the center of the protein, we have named them 'orbit0 = core, orbit1 = inner, orbit2 = middle, and orbit3 = surface'. By ignoring both orbits and polarity specifications or each of them individually, it is possible to define different sub-sets. The calculation of all possible Θ_k (orbit, pol) was carried out using our experimental software MARCH-INSIDE, including the possibility of manual variation of the orbits' frontiers.[23]

In order to illustrate the use of the present approach we have developed a linear discriminant analysis[24] to find a QSAR for 26 proteins with three possible biological activities. The types of proteins considered were lysozymes (L), dihydrofolate reductases (DR), and alcohol dehydrogenases (AD).[25] Classic methodologies based on sequence information, geometric information or physicochemical parameters may perform in a similar way, although we would emphasize that this study is mainly for illustrative purposes, in case a comparison with classic physicochemical parameters is carried out below. The best classification functions we found were:

$$\begin{aligned} \text{L} &= 1.17 \times \Theta_5(\text{core}) + 49.1 \times \Theta_0(\text{middle}) \\ &\quad - 26.5 \times \Theta_3(\text{surface}) + 74.3 \times \Theta_0 - 753.7 \\ \text{AD} &= 1.22 \times \Theta_5(\text{core}) + 53.6 \times \Theta_0(\text{middle}) \\ &\quad - 28.9 \times \Theta_3(\text{surface}) + 82.8 \times \Theta_0 - 930.0 \\ \text{DR} &= 1.09 \times \Theta_5(\text{core}) + 48.1 \times \Theta_0(\text{middle}) \\ &\quad - 24.6 \times \Theta_3(\text{surface}) + 75.9 \times \Theta_0 - 781.8 \quad (4) \end{aligned}$$

It should also be noted that in the present application it was not necessary for considering aminoacid separation with respect to polarity, but only in terms of orbits. These equations correctly classify 92.3%, 24 out of the 26 proteins studied. In particular, the model correctly classifies 88.9% (8 out of 9) of alcohol dehydrogenases, 90% (9 out of 10) of lysozymes, and 100% of the dihydrofolate reductases. The same high classification percentages were also found when testing the predictability of the model using leave-one-out (LOO) procedures. A detailed list of the PDB[26] codes of the 26 proteins together with their observed and predicted classifications, and posterior probabilities are shown in Table 1. The model proved to be statistically significant with a Wilk's lambda (λ_1) of 0.096.

Table 1. PDB codes, observed, predicted classifications, and posterior probabilities for the 26 proteins in the study

Protein	Observed activity	Predicted activity	P (%)	Predicted LOO	P_{LOO} (%)
4LYT A	L	L	99.7	L	99.4
4LYT B	L	L	99.5	L	99.3
1GHL A	L	L	99.8	L	99.7
1GHL B	L	L	99.9	L	99.8
1HNL	L	L	99.1	L	99.1
1LMN	L	L	97.1	L	96.8
1LZ1	L	L	98.5	L	97.7
2EQL	L	DR[a]	6.9	DR[a]	5.9
2IHL	L	L	83.7	L	77.9
135L	L	L	89.1	L	84.7
2OHX A	AD	AD	100.0	AD	99.9
2OHX B	AD	AD	100.0	AD	99.9
1CDD A	AD	AD	96.7	AD	91.7
1CDD B	AD	AD	96.7	AD	91.7
1DEH A	AD	AD	99.9	AD	99.8
1DEH B	AD	AD	99.9	AD	99.8
1QOR A	AD	AD	80.6	AD	87.0
1QOR B	AD	AD	79.8	AD	86.7
1UOK	AD	L[a]	5.6	L[a]	3.0
8DFR	DR	DR	52.8	DR	56.3
1AI9 A	DR	DR	59.4	DR	61.6
1AI9 B	DR	DR	83.8	DR	86.4
1DRF	DR	DR	85.9	DR	88.2
1DYR	DR	DR	78.7	DR	84.1
4DFR	DR	DR	92.4	DR	93.7
3DFR	DR	DR	92.4	DR	93.7

L: lysozymes, AD: alcohol dehydrogenases, DR: dihydrofolate reductases.
[a] Misclassification.

In order to simplify the equations for the purposes of interpretation and the possibility of graphical representation, we performed a canonical analysis.[24] The main root obtained (Root1) proved to be a simple equation centered at 0:

$$\begin{aligned} \text{Root1} &= -0.35 \times \Theta_5(\text{core}) - 2.28 \times \Theta_0(\text{middle}) \\ &\quad + 0.98 \times \Theta_3(\text{surface}) - 1.8 \times \Theta_0 \\ \text{Root2} &= -0.73 \times \Theta_5(\text{core}) - 1.57 \times \Theta_0(\text{middle}) \\ &\quad + 1.49 \times \Theta_3(\text{surface}) + 0.11 \times \Theta_0 \quad (5) \end{aligned}$$

This canonical root presented a high regression coefficient of 0.903, which is statistically significant (p-level<0.05) together with a Chi-squared statistics of 50.33 and a Fisher ratio of 11.12. Overall, the mechanistic interpretation seems to be that the more important step to occur is the protein–substrate interaction that affects all the protein ($-1.8 \times \Theta_0$), but involving mainly aminoacids at the middle orbit ($-2.28 \times \Theta_0$). Subsequently, the effect of this interaction in protein structure is relaxed at topological distance as a result of the propagation of the electrostatic interactions toward the surface ($0.98 \times \Theta_3$), or if necessary (and less probably) inwards the core of the protein ($-0.35 \times \Theta_5$). After direct inspection of the canonical space (Fig. 2) it seems logical to interpret Root1 as an equation that mainly discriminates between Alcohol Dehydrogenases and the remaining proteins. These results are consistent with the position of the catalytic pockets of these proteins

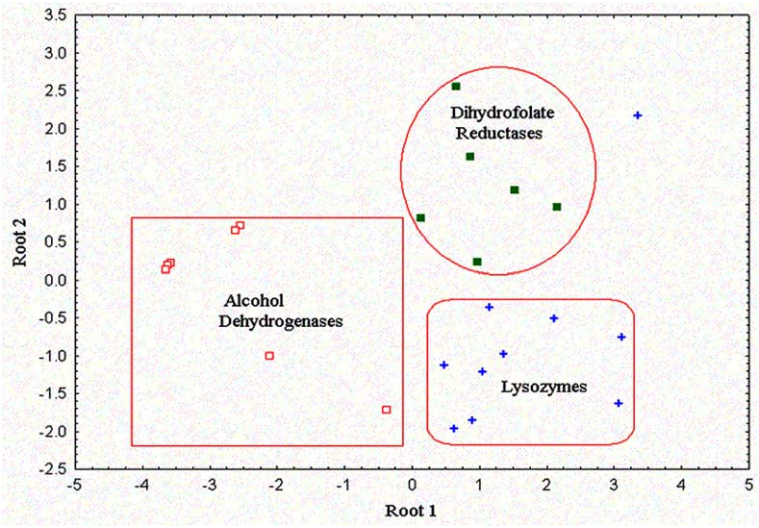

Figure 2. Graphical representation of the 26 proteins within the 2D canonical space.

being distant from the core, but not exactly at the surface of the protein.[25,26]

Conversely, the second root (Root2), which represent a weaker role ($R_{c1} = 0.7$) seems to modulate the differences between lysozymes and dihydrofolate reductases. It is remarkable that in the case of these proteins the surface descriptors maintain their positive contribution but their magnitude changes, meaning that surface changes are more important in order to discriminate between lysozymes and dihydrofolate reductases than for the recognition of alcohol dehydrogenases.

Finally, as a matter of comparison with respect to traditional physicochemical parameters we have calculate the polarizability (Pol), refractivity (M_R), and *n*-octanol/water partition coefficients (P) for all the proteins in this series.[9,27] The best LDA model we found for the present problem, using the same statistical procedure, was:

$$\begin{aligned} L &= 4.04 \times \mathrm{Pol} + 5.57 \times P + 0.53 \times M_R - 2812.8 \\ AD &= 4.28 \times \mathrm{Pol} + 5.81 \times P + 0.52 \times M_R - 3063.2 \\ DR &= 4.07 \times \mathrm{Pol} + 5.63 \times P + 0.54 \times M_R - 2887.5 \end{aligned} \quad (6)$$

This model proved to be statistically significant with $\lambda_2 = 0.194$ but with a relative loss on statistical signification of ~50.5% ($100 \times (\lambda_2 - \lambda_1)/\lambda_2$) relative to the Markov model ($\lambda_1 = 0.096$). In fact, this classic physicochemical model correctly classifies only 80.8% of the proteins in comparison with the 92.3% of accuracy showed by the Markov model. In detail, the model classifies correctly 88.9% (7 out of 9) of alcohol dehydrogenases, 90% (9 out of 10) of Lysozymes, and 71% (2 out of 7) dihydrofolate reductases. These values clearly indicate that although both the Markov and the physicochemical classic models are statistically significant the Markov model is superior in accuracy. Consequently, the present Markov model may perform better than traditional physicochemical parameters in the description of proteins biological activity.

In summary, it has been shown that stochastic molecular descriptors are quite versatile and may be extended to codify the 3D structure of proteins in QSAR. In doing so, it is feasible to model the spherical truncated electrostatic field with an MC. This confirms the previously described importance of spherical truncate electrostatic fields in problems connected with biophysics at the interface between bioorganic and computational chemistry.[28] This paper also corroborates the perspective of QSAR[29,30] techniques, now in the field of proteins.

Acknowledgements

The authors sincerely acknowledge the kind attention of the editor Prof. Dr. D. L. Boger and useful comments from unknown referees. H.G.-D. and E.U. would like to express their gratitude to the Spanish Ministry of Science and Technology (SAF-2003-02222) due to partial financial support.

References and notes

1. Arteca, G. A.; Mezey, P. G. *J. Mol. Graphics* **1990**, *8*, 66.
2. White, J. H. *Am. J. Math.* **1969**, *91*, 693.
3. Flory, P. J. *Principles of Polymer Chemistry*; Cornell University Press: Ithaca, 1953.
4. Fresht, A. *Structure and Mechanism in Protein Science*; W. H. Freeman: New York, 1999.
5. Estrada, E. *Bioinformatics* **2002**, *18*, 1.
6. Estrada, E. *Chem. Phys. Lett.* **2000**, *319*, 713.
7. Cai, Y.-D.; Lina, S. L. *BBA* **2003**, *1648*, 127.
8. Lejon, T.; Strom, B. M.; Svensen, S. J. *J. Pept. Sci.* **2002**, *7*, 74.
9. Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*; Wiley VCH: Weinheim, Germany, 2000.
10. González-Díaz, H.; Olazábal, E.; Castañedo, N.; Hernández, S. I.; Morales, A.; Serrano, H. S.; González, J.; de Ramos, A. R. *J. Mol. Mod.* **2002**, *8*, 237.
11. González-Díaz, H.; Gia, O.; Uriarte, E.; Hernández, I.; Ramos, R.; Chaviano, M.; Seijo, S.; Castillo, J. A.;

Morales, L.; Santana, L.; Akpaloo, D.; Molina, E.; Cruz, M.; Torres, L. A.; Cabrera, M. A. *J. Mol. Mod.* **2003**, *9*, 395.
12. González-Díaz, H.; de Ramos, A. R.; Molina, R. R. *Bull. Math. Biol.* **2003**, *65*, 991.
13. González-Díaz, H.; de Ramos, A. R.; Uriarte, E. *Online J. Bioinf.* **2002**, *1*, 83.
14. González-Díaz, H.; Hernández, S. I.; Uriarte, E.; Santana, L. *Comput. Biol. Chem.* **2003**, *27*, 217.
15. González-Díaz, H.; Marrero, Y.; Hernández, I.; Bastida, I.; Tenorio, E.; Nasco, O.; Uriarte, E.; Castañedo, C. N.; Cabrera-Pérez, M. A.; Aguila, E.; Marrero, O.; Morales, A.; González, M. P. *Chem. Res. Toxicol.* **2003**, *16*, 1318.
16. González-Díaz, H.; Ramos de, A. R.; Molina, R. R. *Bioinformatics* **2003**, *19*, 2079.
17. de Ramos, A. R.; González-Díaz, H.; Molina, R. R.; Uriarte, E. *Proteins: Struct. Func. Bioinf.* **2004**, 10.1002/prot.20159.
18. González-Díaz, H.; Molina, R. R.; Uriarte, E. *Polymers* **2004**, *45*, 3845.
19. González-Díaz, H.; Bastida, I.; Castañedo, N.; Nasco, O.; Olazábal, E.; Morales, A.; Serrano, H. S.; de Ramos, A. R. *Bull. Math. Biol.* **2004**, 10.1016/j.bulm.2003.12.003.
20. Collantes, R. E.; Dunn, J. W., III. *J. Med. Chem.* **1995**, *38*, 2705.
21. Freund, J. A.; Poschel, T. *Stochastic Processes in Physics, Chemistry, and Biology*. In *Lecture Notes in Physics*; Springer-Verlag: Berlin, Germany, 2000.
22. Steinbach, P. J.; Brooks, B. R. *J. Comput. Chem.* **1994**, *15*, 667.
23. González-Díaz, H.; Molina, R. R.; Hernández, I. MARCH-INSIDE®. 2002, version 2.0, (**Mar**kovian **Ch**emicals '**In Si**lico' **De**sign), 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 send an e-mail to the corresponding author: humbertogd@vodafone.es or humbertogd@uclv.edu.cu.
24. Van Waterbeemd, H. Discriminant Analysis for Activity Prediction. In *Method and Principles in Medicinal Chemistry*; Manhnhold, R., Krogsgaard-Larsen, H. T., Eds.; Chemometric Methods in Molecular Design; Van Waterbeemd, H., Ed.; VCH: Weinhiem, 1995; Vol. 2, pp 265–282.
25. Fleming, P. J.; Richards, F. M. *J. Mol. Biol.* **2000**, *299*, 487.
26. Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *NAR* **2000**, *28*, 235.
27. Hypercube Inc. *HyperChem for Windows* 2000, version 6.3.
28. Norberg, J.; Nilsson, L. *Q. Rev. Biophys.* **2003**, *36*, 257.
29. González, M. P.; Morales, A. H.; González-Díaz, H. *Polymer* **2004**, *45*, 2073.
30. González, M. P.; Morales, A. H.; Molina, R. R. *Polymer* **2004**, *45*, 2773.