Tight-Binding "Dihedral Orbitals" Approach to the Degree of Folding of Macromolecular Chains

Ernesto Estrada†

Complex Systems Research Group, X-rays Unit, RIAIDT, Edificio CACTUS, University of Santiago de Compostela, 15706 Santiago de Compostela, Spain

Received: June 14, 2007; In Final Form: September 7, 2007

We develop a tight-binding molecular approach to quantify the degree of folding of a macromolecular chain. This approach is based on the linear combination of "dihedral" orbitals to give molecular orbitals (LCDO-MO). The dihedral orbitals are a set of orbitals situated in each dihedral angle of the chain. The LCDO-MO approach remains basically topological, and we display its direct relation to known graph theoretical concepts. Using this approach, we define the dihedral electronic energy and the dihedral electronic partition function of a linear macromolecular chain. We show that the partition function per dihedral angle quantifies the degree of folding of the dihedral graph. We analyze the empirical relationship between these two functions by using a series of 100 proteins. We also study the relation between these two functions and the percentages of secondary structure for these proteins. Finally, we illustrate the use of the dihedral energy and the partition function in structure−property studies of proteins by analyzing the binding of steroids to DB3 antibody.

Introduction

The classical paradigm of structural biology is to determine the structure of protein in order to understand how it performs its known biological function at the molecular level. This paradigm has been challenged by the current post-genomic research, whose challenge consists of starting from the gene sequence, producing the protein, then determining its three-dimensional structure and finally extracting useful biological information about the biological role of the protein in the organism.[1,2] Currently, there are 46 818 structures deposited in the Protein Data Bank.[3] From these structures, 43 099 correspond to proteins, 1794 correspond to nucleic acids (NA), and 1892 correspond to protein/NA complexes. These proteins are grouped into about 1600 superfamilies and are characterized by between 800 and 1000 unique topologies.[4] Thus, despite the vast array of different functions observed for proteins, they appear to exist in a limited (relatively small) number of families and folds. Because of this tremendous amount of structural data, it is necessary to develop and use new theoretical tools to extract the maximum structural information from protein structures to fulfill the new paradigm of structural genomics. This strategy based on the use of protein descriptors is taking recent relevance for solving the problem of fold assignment,[5,6] as well as for predicting protein−peptide interactions, interactions of proteins with organic molecules, and predicting indirect effects in proteins.[7,8]

One of the most important characteristics of the three-dimensional structure of a protein is its degree of folding (DOF). The necessity for quantifying the DOF of a (macromolecular) chain arises from the continued use of the term "folding" in a quantitative context. It is common to find references in the scientific literature where the expression "A is more folded than B" is used despite no quantitative measure of the DOF of A and B is given.[9] On the other hand, the use of quantitative measures for comparing the DOF of two chains is also frequent despite that the descriptors used do not characterize the DOF of the corresponding chains.[9] The first attempt to assign a quantitative measure to the DOF was carried out by Randić and Krilov[10,11] a few years ago on the basis of the so-called distance/distance matrixes. Balaban and Rücker[12] used "protochirons", which are three-step path conformations. More recently, Liu and Wang[13] have extended this approach by including four new kinds of three-step path conformations for studying the degree of folding of protein chains. All of these methods of quantifying the degree of folding are phenomenological approaches based on the chemical intuition.

A different strategy, which is also a phenomenological approach, was proposed by the present author by using graph spectral theory.[14,15] This index of degree of folding has been very useful in structure−function studies of proteins as well as in protein secondary structure classification. For instance, this index was applied to the study of steroid-antibody binding affinity,[16] for characterizing the contribution of amino acids to the global degree of folding of proteins,[17] and for automating the classification of proteins into their respective structural domain classes, namely, mainly α, mainly β, and $\alpha-\beta$.[18] In other studies, we also found that the reduction potential of azurins and pseudoazurins is related to the contribution of specific amino acids to the global degree of folding[19] and that the degree of folding plays a fundamental role in explaining the binding energetics of protein−ligand interactions, such as protein/peptide and protein/drugs interactions.[20] In the current work, we use simple quantum chemical ideas to define the degree of folding of a (macromolecular) chain. This approach is based on a tight-binding Hamiltonian based on orbitals centered at the dihedral angles of the linear chain. We show that the "electronic dihedral energy" reflect important aspects of the folding of a protein. At the same time, we show that the "electronic dihedral" partition function is intimately related to the spectral measure of degree of folding previously proposed by the present author.[14,15] Finally, we use both measures for

† Corresponding author. Fax: 34 981 547 077. E-mail address: estrada66@yahoo.com.

10.1021/jp074595x CCC: $37.00
Published on Web 11/08/2007

studying the degree of folding of 100 protein chains as well as to study structure−function relationships in proteins.

Theoretical Antecedents

First, we introduce here the spectral measure of degree of folding that we have previously defined,[14,15] which is the motivational idea of the current work.

The geometry of a linear molecular chain is determined by the bond distances, the bond angles, and the dihedral angles formed between the atoms of the chain. It is well-known that both the bond distances and the bond angles remain practically unaltered when a molecule changes from one conformation to another. In such changes, the geometrical parameter that makes the difference is the dihedral angle.

Let us first consider a chain formed by four atoms numbered in successive order as 1, 2, 3, and 4. The dihedral angle is defined as the angle between the plane formed by the atoms 1,2,3 and the plane formed by the atoms 2,3,4. In general, any dihedral angle *i*,*j*,*k*,*l* is formed by the intersection of the two planes *i*,*j*,*k* and *j*,*k*,*l*. These chains are known as polygonal chains or discrete curves.[21,22] Thus, it is natural to think about the dihedral angle when we want to compare the DOF of two different chains. In the case of the chain formed by only four atoms, the closer to zero the dihedral angle is, the larger the degree of folding of the chain is. At the other extreme, the extended chain in which the dihedral angle is equal to 180° is the least folded conformation.

However, when we try to extend this intuition to a linear chain having more than one dihedral angle, we realize that there is another factor that influences the degree of folding of the chain. This second factor is the "distribution" of the dihedral angles along the chain. For the sake of simplicity, let us next consider a linear chain having three dihedral angles. Then, starting from the extended conformation (180°, 180°, 180°), we can obtain two conformations having a folded angle, for example, 0°, 180°, 180° and 180°, 0°, 180°, as well as two conformations with two folded dihedrals, for example, 0°, 180°, 0° and 0°, 0°, 180°. This begs the question which of the conformers in each pair represents the most folded chain. We intuitively think that the chain is folded most when a kink is positioned at the center. Thus, the 180°, 0°, 180° conformation is considered to be more folded than 0°, 180°, 180°. On the other hand, we can assume that the closer the kinks are to each other, the more folded the chain is. Consequently, the 0°, 0°, 180° conformation is more folded than 0°, 180°, 0°. Following these intuitive rules for longer chains spells a very difficult and error prone process. Thence, we have devised a mathematical approach for accounting quantitatively the above-mentioned intuitive reasoning.

Using the above-mentioned intuitions, we represent a linear chain by means of the adjacency between the dihedral angles of the chain, that is, a dihedral chain. Thus, for a chain having Q atoms, there are $N = Q - 3$ dihedral angles, which are represented by the nodes of the dihedral graph (see Figure 1). The dihedral angles represented in this figure are designated as usual for protein chains with the symbols ϕ, ψ, and ω. The angle ω_i defines the rotation about the C_i-N_{i+1} peptide bond. ϕ_i describes the rotation about the $N_i-C\alpha_i$ bond, and ψ_i describes the rotation about the $C\alpha_i-C_i$ bond. Two nodes in the dihedral chain are connected if the corresponding dihedral angles have a bond in common, that is, they are adjacent. Then, we assign to each node of the dihedral chain a function of the dihedral angles which takes the maximal value when the angle is equal to zero (most folded conformer) and decreases as the angle increases.

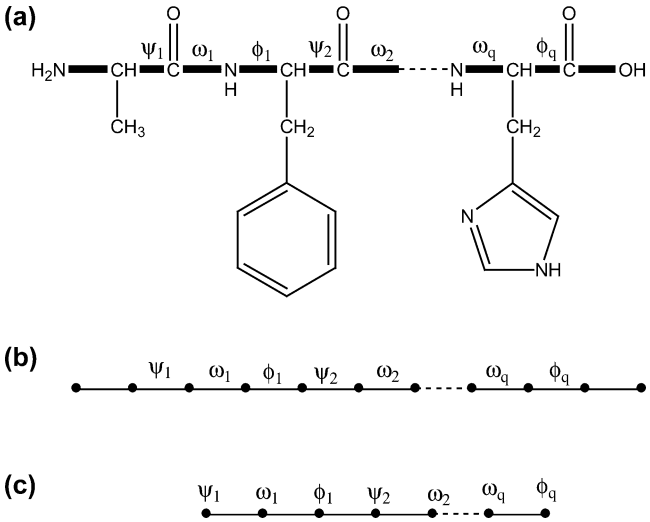

Figure 1. Illustration of a protein structure (a), its backbone chain (b), and the dihedral graph (c) obtained for it.

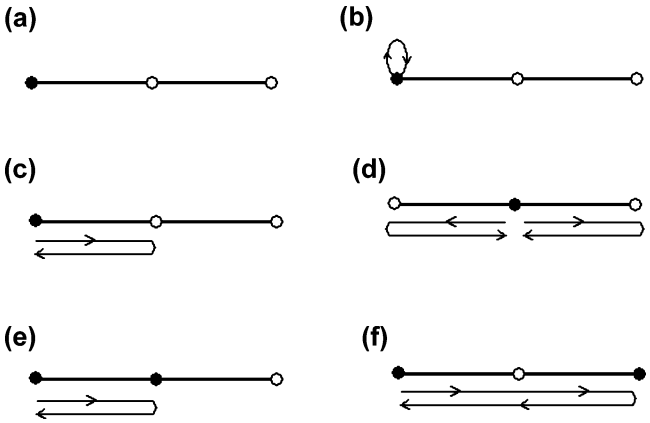

Figure 2. Illustration of closed walks (CWs) of different lengths. (a,b) CWs of length zero and one, respectively. (c,d) CWs of length two which are able to differentiate between two folded angles (marked as black). (e,f) CWs of length 2 and 4, respectively, which visit two folded dihedrals separated at different distances.

The adjacency relationships between the dihedral angles of the chain can be represented through the use of the adjacency matrix of the dihedral chain **A**. The non-diagonal entries of this matrix are 1's or 0's if the corresponding dihedral angles are adjacent or not, respectively. In order to account for the dihedral angle function, we use the diagonal matrix $\mathbf{V} = \mathrm{diag}(V_1, V_2, ..., V_N)$, where V_i is the function of the ith dihedral angle. Finally, we consider the weighted adjacency matrix $\mathbf{W} = \mathbf{A} + \mathbf{V}$ in order to describe the dihedral chain.

Our strategy to measure the degree of folding of the chain is to count the number of walks which starts and ends on a node of the dihedral chain, that is, closed walks. For instance, the closed walks (CWs) of length 0 correspond to simply staying at the node. Then, the sum of all CWs of length 0 in the dihedral chain is the number of dihedral angles in the chain. A CW of length 1 corresponds to going out and returning to the corresponding node by using the loop attached to it (see Figure 2). Then the sum of CWs of length 1 is equal to the sum of the functions assigned to the dihedral angles. Using this second measure, we can differentiate a conformation having one kink from an extended conformation.

However, in order to differentiate two conformations having one kink at different positions, we need to count the number of CWs of length 2. A CW of length 2 is the one that goes from

one node to its neighbor and returns. Then, in the example given below where the kink was at the center of the chain, there will be two CWs of length 2 involving this dihedral angle while there is only one CW of length 2 involving the kink which is at the end of the chain. We can continue by extending the length of the CWs to account for all other distributions of dihedral angles along the chain. For instance, if we want to differentiate the conformations 0°, 180°, 0° and 0°, 0°, 180°, we have to see that in the first conformer if we want to go from one kink to the other and return to the first we need to use a CW of length 4 (see Figure 2). However, in the second chain conformation, we only need a CW of length 2 for doing the same. Consequently, it is necessary to give different weights to CWs of different lengths. CWs of shortest lengths should receive larger weights when we sum them to account for the degree of folding.

The number of CWs of length k starting (and ending) at node i in the chain is determined by the ith diagonal entry of the kth power of the weighted adjacency matrix $\mathbf{W}$ of the chain, $(\mathbf{W}^k)_{ii}$.[23] Then, as a measure of the degree of folding of the chain, we proposed to use the sum of the number of CWs of different lengths by dividing each CW by the factorial of its length:[14,15]

$$I_3 = \frac{1}{N}\sum_{k=0}^{\infty}\sum_{i=1}^{N}\frac{(\mathbf{W}^k)_{ii}}{k!} \tag{1}$$

By this way, we give larger weights to the shortest CW, and we also take advantage of the graph spectral theory[23] in order to calculate the I_3 index on the basis of the eigenvalues λ_j of the adjacency matrix of the dihedral chain:[14,15]

$$I_3 = \frac{1}{N}\sum_{k=0}^{\infty}\sum_{i=1}^{N}\frac{(\mathbf{W}^k)_{ii}}{k!} = \frac{1}{N}\sum_{j=1}^{n} e^{\lambda_j} \tag{2}$$

The index I_3 is the spectral measure of the degree of folding of a chain which was previously defined by the present author[14,15] and has been successfully applied to study the degree of folding of protein chains.[16−20]

Results and Discussion

Tight-Binding "Dihedral Orbitals" Approach. Now, we develop the theoretical principles of the approach that we are introducing in the current work. Our strategy here is that, instead of using atomic-centered orbitals as the basis of the molecular wave functions, we consider a set of orbitals situated in each dihedral angle of the chain. This approach is similar to the one developed by Lennard-Jones and Hall, who defined "equivalent orbitals" to be localized orbitals formed between two bonded atoms.[24−27] Here, we do not go into detail of how these "dihedral" orbitals are formed quantitatively. Instead, we give a qualitative picture of this intuition.

By using atomic orbitals, a set of equivalent orbitals centered on each bond can be built. For instance, Pauling constructed such orbitals as linear combinations of atomic orbitals s, p_z, d_z, and f_z with the maximum values along the z axis, which is taken as the bond direction.[28] Schematically, we illustrate in Figure 3a such atomic orbitals in the direction of the bonds and the corresponding equivalent bond orbitals obtained from them (Figure 3b). The bond orbitals are represented by means of the so-called line graph of the chain. The line graph $L(G)$ is the graph in which the bonds of the chain G are represented as the nodes of $L(G)$. Two nodes of $L(G)$ are adjacent if the corresponding bonds in G share an atom.[29] We can extend this approach to consider the second line graph $L^2(G)$ of the

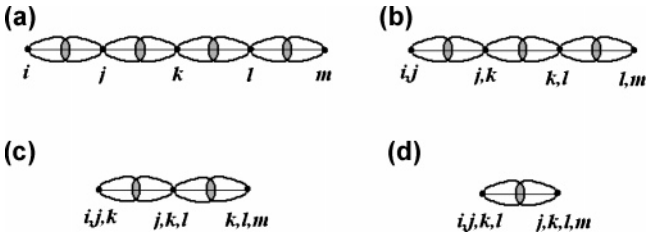

Figure 3. Illustration of the iterated building of orbitals as proposed in the current work: atomic (a), bond (b), plane (c), and dihedral (d) orbitals. The overlapping of two orbitals are shadowed in gray.

molecular chain in which every node represents a bond angle. This approach is known as the iterated line graph sequence, and it has been used in the study of organic molecules and macromolecules.[20−32] The corresponding orbitals located on the plane formed by this angle will be designated as the plane orbital (see Figure 3c). Finally, we consider the existence of "dihedral" orbitals, which are localized between two planes formed by the angles i,j,k and j,k,l as illustrated in Figure 3d.

The molecular orbitals for the dihedral-electrons in the atomic chain can be written as

$$|\Psi_n^D\rangle = \sum_{i=1}^{N} C_n^D(i)|\phi_i^D\rangle \tag{3}$$

where ϕ_i^D is a united-atom orbital located in the ith dihedral bond of the chain. Equation 3 represents the basis of the linear combination of dihedral orbitals to give molecular orbitals (LCDO-MO), similar to the linear combination of bond orbitals to give molecular orbitals developed by Brown.[33] Then, we assume that the "electronic dihedral energy" of a chain can be obtained by solving a dihedral version of the Schrödinger equation:

$$\mathbf{H}^{\mathrm{D}}|\Psi^{\mathrm{D}}\rangle = \epsilon^{\mathrm{D}}|\Psi^{\mathrm{D}}\rangle \tag{4}$$

where the superscript D is used to designate the dihedral angles, and the Hamiltonian is obtained as $\mathbf{H}^{\mathrm{D}} = \mathbf{A} - \mathbf{V}$, where $\mathbf{A}$ is the adjacency matrix of the dihedral angles chains and $\mathbf{V}$ is a dihedral angle function. This Hamiltonian as well as the dihedral angle potential will be considered in more detail further in this work. Substituting eq 3 into eq 4, we have

$$\sum_i C_n^D(i)\mathbf{H}^{\mathrm{D}}|\phi_i^{\mathrm{D}}\rangle = \epsilon^{\mathrm{D}}\sum_i C_n^{\mathrm{D}}(i)|\phi_i^{\mathrm{D}}\rangle \tag{5}$$

Multiplying both sides of this expression from the left by $\langle\phi_j|$ and moving the right-hand side to the left-hand side yields the secular equation, where we have removed, for the sake of simplicity, the superindex D in the dihedral Hamiltonian:

$$\sum_i C_{ni}(H_{ji} - \epsilon S_{ji}) = 0 \tag{6}$$

where $H_{ji} = \langle\phi_j|\mathbf{H}|\phi_i\rangle$ and $\mathbf{S}_{ji} = \langle\phi_j|\phi_i\rangle$. The nontrivial solutions of eq 6 are obtained by solving the determinant equation

$$|\mathbf{H} - \epsilon\mathbf{S}| = 0 \tag{7}$$

We assume that the Coulomb integral $\mathbf{H}_{ii}$ of a dihedral orbital ϕ_i depends only on the angle between the two planes forming the dihedral orbital. The resonance integral $\mathbf{H}_{ij}$ between dihedral orbitals ϕ_i and ϕ_j is assumed to be 0, unless i and j are adjacent dihedrals in the chain, in which case it is taken to be $H_{ij} = q$. We consider that the dihedral orbitals are orthonormal; thus, $S_{ij} = \delta_{ij}$. The Coulomb integrals are set equal to $H_{ii} = p - V_i q$,

where $V_i = f(\varphi_i)$. Here, φ_i is the ith dihedral angle of the linear chain. We use the Coulomb integrals as $H_{ii} = p - V_i q$ in order to obtain the minimum dihedral energy for the most folded conformer as will be evident later.[34] Then, if we divide every entry of the secular determinant by q, the secular determinant for a linear chain having N dihedral angles is written as

$$\begin{vmatrix} \mu - V_1 & 1 & 0 & 0 & \cdots & 0 & 0 \\ 1 & \mu - V_2 & 1 & \ddots & \ddots & 0 & 0 \\ 0 & 1 & \mu - V_3 & \ddots & \ddots & 0 & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \mu - V_{N-1} & 1 \\ 0 & 0 & \cdots & \cdots & \cdots & 1 & \mu - V_N \end{vmatrix} = 0 \quad (8)$$

where $\mu = (p - \epsilon)/q$ are the eigenvalues of the matrix $\mathbf{H} = \mathbf{A} - \mathbf{V}$. This matrix is a tri-diagonal matrix $\mathbf{H} = [H_{ij}]$, which means that $H_{ij} = 0$ whenever $|i - j| > 1$.[35]

The orbital energy is determined by the eigenvalues μ_j of a topological matrix representing the adjacency between pairs of dihedral angles in the chain:

$$\epsilon_j = p - q\mu_j \quad (9)$$

Here, all dihedral orbitals happen to be fully occupied, and the total electronic "dihedral" energy (dihedral energy for brief) is given by[36]

$$E_{\text{dih}} = \sum_{j=1}^{N} 2\epsilon_j = Np - 2q\sum_{j=1}^{N} \mu_j \quad (10)$$

where N is the number of dihedral angles in the linear chain and $q < 0$. From now on, we set $p \equiv 0$ which makes

$$E_{\text{dih}} = 2|q|\sum_{j=1}^{N} \mu_j \quad (11)$$

Electronic "Dihedral" Partition Function. A molecule having N dihedral angles has N topological states for every conformation. In order to understand this interpretation, we need to clarify what is the meaning of "states" in this case. To start with, we consider two different conformations of a linear chain with three dihedral angles. Because these two conformers have three dihedral angles, they have three different microstates. Each of these microstates can be visualized as amplitudes of harmonic waves at the dihedral angle positions. Such amplitudes are given by the coefficients of the corresponding eigenvectors of the dihedral Hamiltonian matrix. The square of the coefficients of the corresponding eigenvectors give the probability of finding the dihedral electron at the corresponding dihedral angle. To summarize, every microstate represents the probability pattern of finding these electrons at the different sites of the dihedral-angle graph. The energy of each of these microstates is given by ϵ_j as expressed by eq 9.

Now, suppose that the "strength" of interaction between any pair of dihedral angles is equal to β. Then the partition function[37] of the dihedral-angle graph is given by

$$Z(\beta) = \sum_{j=1}^{n} e^{-\beta|q|\epsilon_j} = \sum_{j=1}^{n} e^{-\beta|q|\mu_j} \quad (12)$$

where $\beta = 1/kT$ is the inverse temperature and k is the Boltzmann constant.[37] Further, in this work, we consider $|q|\beta \equiv 1$ in all calculations of the partition function, where $|q|$ specifies an energy scale chosen arbitrarily. Thus,

$$Z(|q|\beta = 1) = Z = \sum_{j=1}^{N} e^{-\mu_j} \quad (13)$$

Now, because the matrixes $\mathbf{W} = \mathbf{A} + \mathbf{V}$ and $\mathbf{H} = \mathbf{A} - \mathbf{V}$ are tri-diagonal matrixes, it is easy to prove that eig($\mathbf{W}$) = −eig($\mathbf{H}$), which in other words means that $\lambda_j \equiv -\mu_j$. The immediate consequences of this equality are that

$$E_{\text{dih}} = -2\sum_{j=1}^{N} \lambda_j \quad (14)$$

$$Z = \sum_{j=1}^{N} e^{-\mu_j} = \sum_{j=1}^{N} e^{\lambda_j} \quad (15)$$

By using this partition function, it is possible to define thermodynamic functions of the dihedral-electronic states of the protein chain. These functions will not be considered here for the time being.

An important difference between these two functions, the dihedral energy and the partition function, is the following. The dihedral energy is calculated as the sum of all eigenvalues of the weighted adjacency matrix of the dihedral chain. Then, it is straightforward to realize that it is equal to the sum of the diagonal entries of this matrix,

$$E_{\text{dih}} = -2\sum_{j=1}^{N} V_j \quad (16)$$

This means that the dihedral energy function accounts only for the amount of degree of folding contained in the corresponding dihedral angles of the chain but not on their distribution along the chain. In other words, if there are two chains having the same sum of the dihedral angle functions, they will display the same dihedral energy despite that they can display a different distribution of such angles in the chain. A simple example is given by any of the pairs of chains analyzed previously in this work, like 0°, 180°, 180° and 180°, 0°, 180°. However, the partition function accounts for both the amount of degree of folding contained in the corresponding dihedral angles of the chain and their distribution along the chain.

Topological Connection. An important characteristic of the current approach is that it remains essentially topological. For instance, using a Taylor series, we can expand the partition function (eq 12) to obtain an expression based on the sum of the diagonal entries of the weighted adjacency matrix,

$$Z(\beta) = \sum_{j=1}^{n} e^{\beta\lambda_j} = \sum_{k=0}^{\infty}\sum_{j=1}^{N} \frac{\beta^k(\lambda_j)^k}{k!} = \sum_{k=0}^{\infty}\sum_{j=1}^{N} \frac{\beta^k(\mathbf{W}^k)_{jj}}{k!} \quad (17)$$

The immediate consequence of this expression is that we can consider the temperature as a topological parameter. If we see the right part of eq 17, we can see that β can be considered as a parameter which is multiplying each entry of the weighted adjacency matrix. In other words, if the dihedral angles i and j are adjacent, the i,j entry of $\mathbf{W}$ is equal to β. In addition, the ith diagonal entries of $\mathbf{W}$ are now equal to $\beta(p - qV_1)$. Consequently, the inverse temperature β represents the "strength" of the interaction between dihedral angles as well as their self-interaction.[38] At very large temperature, $\beta \approx 0$ which makes all entries of $\mathbf{W}$ equal to 0. This situation is similar to the

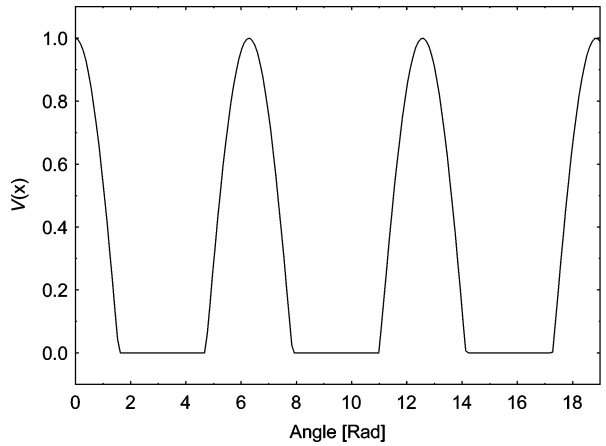

Figure 4. Representation of the half-cosine function used in the Coulomb integrals of the current approach.

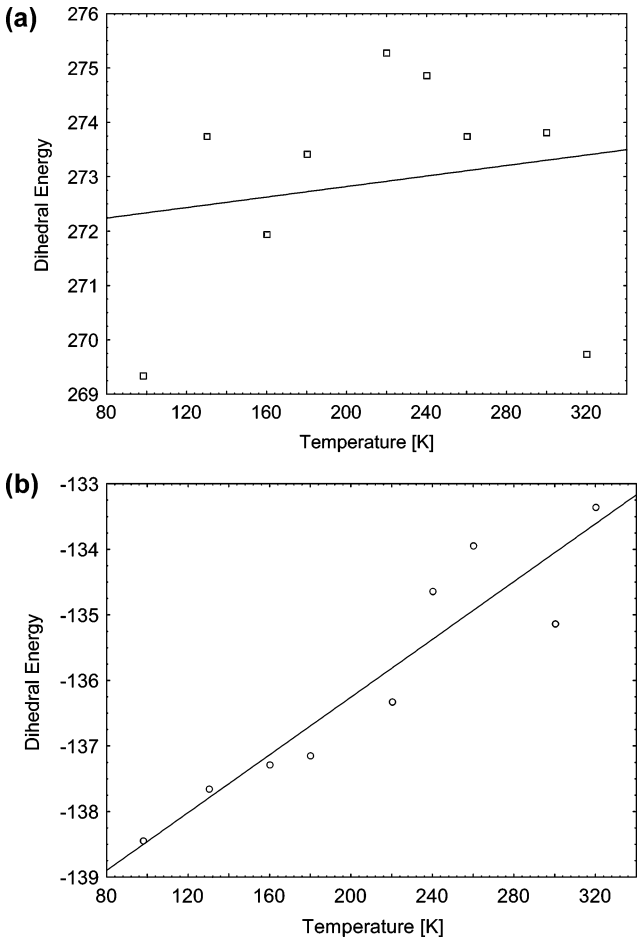

Figure 5. Plot of the dihedral energy of RNAse A (in units of β) as a function of the temperature at which the structure was determined by X-rays. In (a) we have used a cosine function (16), and in (b), we have used a half-cosine function (17) in the Coulomb integrals of the tight-binding "dihedral orbitals" approach.

consideration of the chain as an isolated series of nodes without any interaction between each other.

This finding has an important consequence for the consideration of the degree of folding of macromolecular chains. It means that the folding degree index is the electronic dihedral partition function of the chain, $I_3 = (Z(|q|\beta = 1)/N)$. Generalizing this index for any temperature, we can say that I_3 gives an indication of the average number of states that are thermally accessible to a conformer at the temperature of the system. At $T = 0$, only the ground level is accessible. At very high temperatures, virtually all states are accessible.

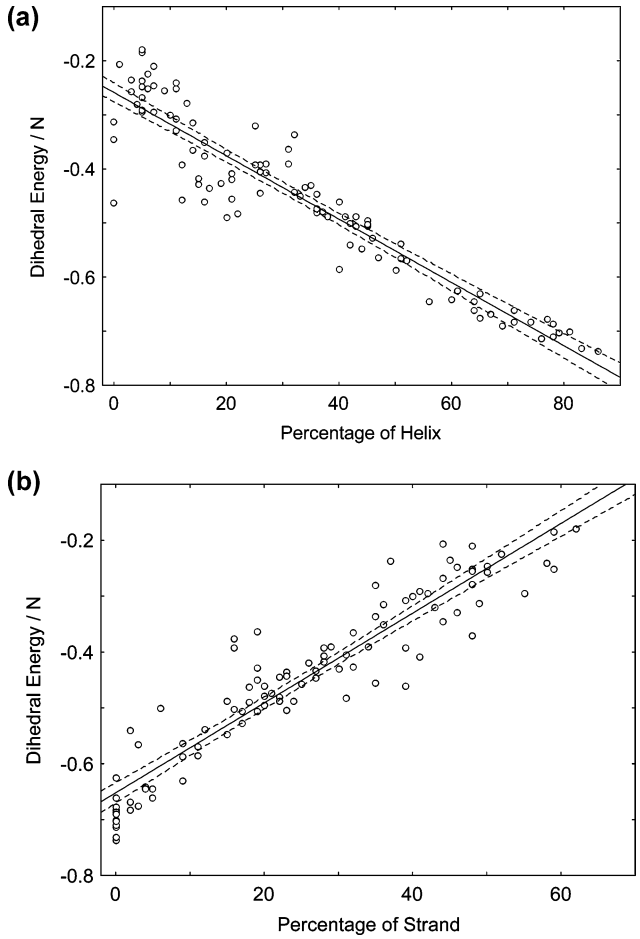

Figure 6. Linear correlations between the dihedral energy per dihedral angle versus the percentage of helix (top) and the percentage of strand (bottom) for 100 proteins studied in the current work.

Dihedral Angle Function. An important aspect of the current approach is the appropriate selection of the function V_i which appears in the Coulomb integrals. We consider here that this function depends only on the angle φ_i formed between the two planes determining the dihedral. According to our intuitive model sketched in Figure 3, we find the maximum overlapping between the plane orbitals is obtained when the angle between these two planes is equal to 0°. As this angle increases, the overlapping between the plane orbitals decrease and consequently the function V_i should also decrease. Then a natural way of selecting this function is to make it equal to the cosine of the dihedral angle,

$$V_i = \cos \varphi_i \tag{18}$$

Accordingly, $-1 \leq V_i \leq 1$ taking the minimum for the dihedral angle $\varphi_i = 180°$. According to our intuition when $\varphi_i = 90°$, there is no overlapping between the plane angles forming the dihedral, and V_i should be equal to 0. This condition is fulfilled by the previous function $V_i = \cos \varphi_i$. However, the question that naturally arises here is whether V_i should take negative values (as indicated by $V_i = \cos \varphi_i$) for $90° < \varphi_i \leq 180°$ or simply it should be equal to 0, indicating no overlap at all between the plane orbitals as our intuition indicates. In order to test empirically these two

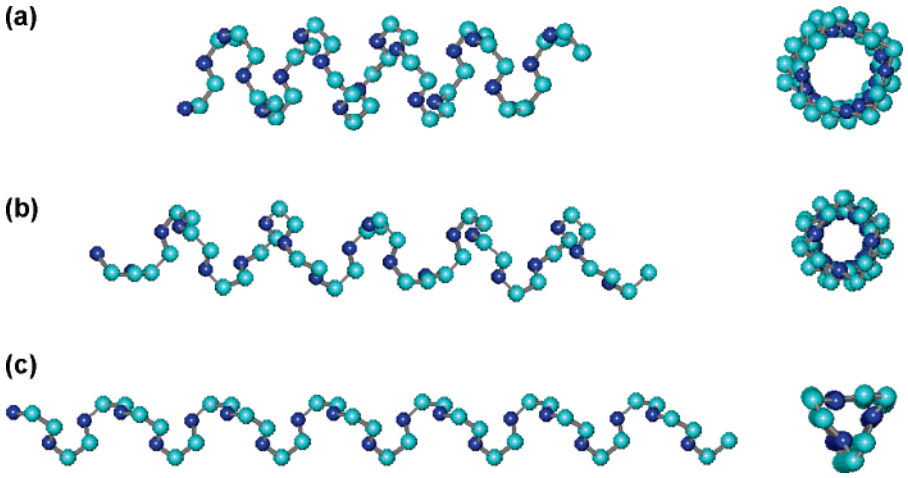

Figure 7. Lateral (left) and top (right) views of a peptide chain in three different helix conformations: π helix (a), α helix (b), and 3_{10} helix (c).

possibilities, we select another form for the function V_i which is given as follows:[39]

$$V_i = \frac{1}{2}[1 + \mathrm{sgn}(\cos \varphi_i)]\cos \varphi_i \qquad (19)$$

where sgn(x) is the sign function defined as

$$\mathrm{sgn}(x) = \begin{cases} -1, & x < 0 \\ 0, & x = 0 \\ 1, & x > 0 \end{cases} \qquad (20)$$

Then this function takes values equal to cosine of the angle for $0° \leq \varphi_i \leq 90°$ and values equal to 0 for $90° < \varphi_i \leq 180°$ as illustrated in the plot of V_i given in Figure 4.

In order to test the two cosine functions used for the Coulomb integrals in the current approach, we selected a series of proteins which were studied experimentally by X-rays at nine different temperatures.[40] The proteins in question correspond to the RNAse A and have PDB[3] codes 1rat, 2rat, 3rat, 4rat, 5rat, 6rat, 7rat, 8rat, and 9rat.

First, we calculated the dihedral energy using both functions eqs 18 and 19 for the backbones of the RNAse at temperatures from 98 to 320 K. It is expected that at low temperatures the protein chains are more folded. As a consequence, the protein is expected to have lower dihedral energy at these low temperatures. As the temperature increases, the degree of folding decreases and the dihedral energy should increase. In other words, it is expected that the dihedral energy displays a positive linear correlation with the temperature. In Figure 5, we illustrate the correlations obtained between the temperature and the dihedral energy calculated by using the cosine function (a) and the half-cosine function (b). The first has a poor correlation coefficient of 0.17 while the second displays a good correlation coefficient of 0.93. In a similar way, when we calculate the I_3 index using the cosine function, we obtain the correlation coefficients of -0.688 contrasting with the value of -0.921 obtained when using the rectified half-wave cosine. The conclusion of this phenomenological experiment is that the use of the half-cosine function is more appropriate than the simple cosine function to be used in the Coulomb integrals of the current approach. This selection completely coincides with our previous intuitive definition of the term to be accounted for the Coulomb integral in the tight-binding dihedral orbitals approach.

Electronic Dihedral Energy of Proteins Chains. Here, we are mainly interested in investigating how much duplicated information the dihedral energy and partition function contain and how they are related to the percentage of secondary structure in proteins. We have selected 100 protein chains whose structures have been reported by using X-rays diffraction. The Protein Data Bank (PDB)[3] codes for these proteins are the following: 3sgb, 8rxn, 1knt, 1pgb, 1rop, 1tgs, 7pti, 1isu, 3ebx, 1cse, 1ptx, 2sn3, 3il8, 1hoe, 1fia, 1pk4, 2bop, 2hpe, 1cmb, 9rnt, 1aaj, 1fkb, 1cew, 1rtp, 1ccr, 2msb, 2tgi, 2hmz, 2rsp, 2chs, 1srg, 1etb, 1poa, 1pmy, 7rsa, 2ccy, 3chy, 2aza, 2ihl, 1ifc, 1lis, 1mdc, 1rsy, 1eca, 1kab, 2end, 2fox, 1nhk, 3sdh, 1bab, 2hbg, 1cob, 2gdm, 2mge, 2rn2, 1hfc, 1gpr, 119l, 1cpc, 2cpl, 1huw, 5p21, 1ofv, 1lki, 1bbp, 1erb, 2scp, 4gcr, 1ytb, 1len, 153l, 1lts, 1rec, 1xnb, 1gky, 1knb, 1dsb, 1isc, 1cus, 2alp, 1iae, 1iag, 1sac, 1scfb, 1thv, 2abk, 1ppn, 3cla, 2ayh, 8fab, 2gst, 1hne, 1dts, 2ak3, 1hsl, 1rva, 1tph, 1mrg, 1nba, 4blm.

We calculate the dihedral energy, E, and the partition function, $Z \equiv Z(|q|\beta = 1)$, for the backbones of these proteins using the function, eq 19. Then, we have correlated E versus Z for the complete series of proteins studied, and we observe that both measures display a modest linear correlation with correlation coefficient $r = -0.86$. This result indicates that one of the measures is able to explain 74% of the variance observed in the other. We have previously shown that the index of degree of folding I_3 is linearly related to the percentages of helix and strand in the proteins. We investigate here the relationships of the dihedral energy and the partition function with the number of dihedral angles (N), the percentage of helix (H), and the percentage of strand (S) in the protein. The partition function is strongly correlated to the size of the chain (N) displaying a correlation coefficient of 0.975. However, the dihedral energy is less correlated to the chain size, having a modest correlation coefficient of -0.726. These correlations are expected from the fact that both properties are extensive; that is, they depend on the size of the systems. When E and Z are expressed in terms of the percentage of secondary structure, H and S, the results display the dissimilarities between these two measures. While E displays modest linear correlation with H and S ($r = 0.715$), Z has no correlation at all with these parameters ($r = 0.347$). However, after the normalization of both parameters by dividing them by N, the correlation coefficients increase dramatically up to -0.957 and -0.966, respectively. This indicates that the degree of folding is better described by using the normalized indices, that is, E/N and I_3, indicating the intensive character of the degree of folding. The parameter E/N represents the dihedral energy per dihedral angle in the chain. The plots of E/N versus

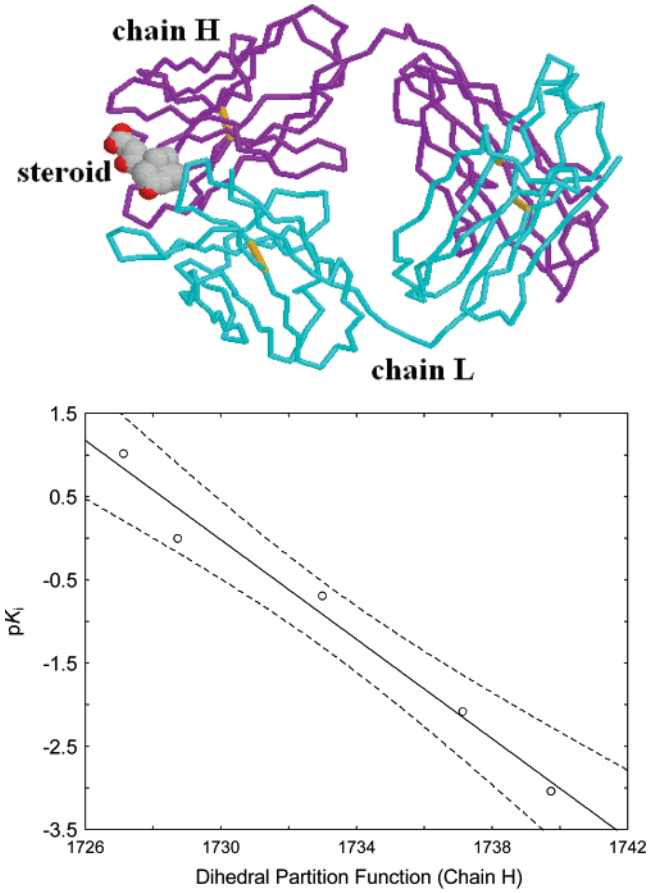

Figure 8. Illustration of the structure of the DB3-monoclonal antibody interacting with a steroid (top). The heavy (H) chain of the antibody is in magenta, and the light chain is in cyan. (bottom) Linear correlation between the logarithm of the binding constant of steroids to DB3 and the dihedral partition function of the H chain of the steroid, which measures the degree of folding of this chain.

the percentages of secondary structure in these proteins are illustrated in Figure 6.

The existence of these linear correlations between the dihedral energy and the partition function with the percentage of secondary structure does not mean that both kinds of measures contain exactly the same structural information. In fact, the dihedral energy and partition function are able to differentiate structures containing the same percentage of secondary structure. For instance, in the previous example of the RNAse A studied at different temperatures, the protein maintains its percentages of helix and strands invariable at the different temperatures studied. However, we have seen that E and Z vary systematically with the temperature. In addition, we illustrate in Figure 7 a peptide chain of 20 amino acids in 3 different conformations, which are 100% helix. They correspond to a π helix, an α helix, and a 3_{10} helix. It is known that the 3_{10} helix is the most folded one and that the π helix is the least folded one. The values of E and Z for these three chains are $E(\pi) = -33.693$, $Z(\pi) = 175.388$, $E(\alpha) = -46.053$, $Z(\alpha) = 197.554$, $E(3_{10}) = -59.085$, $Z(3_{10}) = 225.606$. As can be seen, both measures give different values for the three different chains. E and Z clearly indicate that the order of degree of folding for these chains is $\pi < \alpha < 3_{10}$, which agrees with the chemical intuition.

Application to Structure−Function Relationships Studies. As a final example of the possibilities of the dihedral energy and the partition function, we study the binding interaction of five steroids with the DB3 antibody.[41] The DB3 antibody is produced by the immunization of mice with 11-α-progesterone hemisuccinate coupled to bovine serum albumin. It has been shown that this antibody displays cross-reactivity to a set of structurally diverse steroids at nanomolar concentrations. The X-rays crystallographic structures of five of these complexes were reported by Arevalo et al.[41] We have used these structures to calculate the dihedral energy and the partition function of the antibody chains. Antibodies have two chains designated as heavy (H) and light (L) chains as illustrated at the top of Figure 8. Then, we obtain linear structure−function relationships between the logarithm of the inverse binding affinities, $pK_i = -\ln K_i$ (in nM) with E and Z for each chain (H and L). The steroids bound to the DB3 antibody and their pK_i are as follows:[41] progesterone-11α-ol-hemisuccinate (1.022), progesterone (0.000), 5α-pregnane-3β-ol-hemisuccinate (−0.693), aetiocholanone (−3.044), and 5β-androstane-3,17-dione (−2.079). While the energy and the partition function of the chain L give correlations explaining only 80% of the variance of the pK_i values, these parameters for the chain H explain more than 95% of such variance. The linear regression models obtained for the chain H are given below together with the statistical parameters of the regression:

$$pK_i = 74.627 + 0.492E$$

$$n = 5 \quad r = 0.975 \quad s = 0.404 \quad F = 61.4 \qquad (21)$$

$$pK_i = 516.71 - 0.812Z$$

$$n = 5 \quad r = 0.989 \quad s = 0.276 \quad F = 135.6 \qquad (22)$$

where r is the correlation coefficient, s is the standard deviation, and F is the Fisher ratio of the regression model. As can be seen, both theoretical parameters explain very well the variance in the experimental values of the binding affinities of steroids to DB3 antibody. In particular, the dihedral partition function explains almost 98% of the variance in the pK_i values. The corresponding linear correlation is illustrated at the bottom of Figure 8. The slightly better correlation obtained by using the partition function may indicate that not only the degree of folding of the dihedral angles in the protein but also their distribution along the chain are important to describe the energetic of this interaction.

We have also calculated the values of E and Z for the DB3 antibody without steroids, which are −150.186 and 1726.615, respectively. These values indicate that after the binding of steroids the degree of folding of the chain H of the DB3 antibody increases. This increment of the degree of folding is translated into a stabilization of the chain according to its dihedral energy. The results obtained here with the dihedral energy and the dihedral partition function are significantly better than those obtained by Chen et al.[42] using a linear interaction energy/molecular dynamics to calculate the energy of binding of these steroids to DB3 antibody.

Conclusions

We have introduced in this work a tight-binding approach based on a linear combination of dihedral orbitals to give molecular orbitals (LCDO-MO) to study the degree of folding of (protein) chains. Using this approach, we have defined the dihedral electronic energy for a linear macromolecular chain, like the protein or nucleic acids backbones. This dihedral energy is calculated from the topological adjacency matrix of a graph representing the adjacency between dihedral angles in the chain by using appropriate weights in the main diagonal. The topological matrix represents the Hamiltonian of the LCDO-

MO approach, and the diagonal weights are the corresponding potentials for the Coulomb integrals. We have also defined the dihedral electronic partition function for the macromolecular chains, which represents the degree of folding of the chain. Thus, the degree of folding gives an indication of the average number of states that are thermally accessible to a macromolecular chain at the temperature of the system.

Using the dihedral energy and the partition function, we have studied the influence of the temperature on the degree of folding of the chain of RNAse A. We have seen that as the temperature increases the partition function measuring the degree of folding systematically decreases. This de-folding of the chain is traduced in an increment of the dihedral energy of the chain. In general, we have shown that the dihedral energy per dihedral angle decreases as the percentage of helix in the proteins increases. At the same time, the dihedral energy increases as the proteins contain a higher percentage of strands. The dihedral energy and the partition function are sensible not only to the changes of the temperature and percentage of secondary structure but also to the changes of helix types, for example, π, α, and 3_{10} helices. We have also shown that the current approach is useful to study structure−property relationships in proteins by analyzing the binding of steroids to DB3 antibody. In closing, we believe that the LCDO-MO approach represents an interesting alternative to study the degree of folding of proteins and nucleic acid chains. In further works, we will study other properties derived from this approach, such as the electronic communicability through macromolecular chains.

Acknowledgment. This work is dedicated to Professor George G. Hall for his pioneering and relevant contributions to quantum mechanics and graph theory in chemistry.The author thanks N. Hatano (University Tokyo) and Y. Simón-Manso (NIST, Maryland) for interesting and useful suggestions. Encouraging words from Professor G. G. Hall were a great support for this work. The "Ramón y Cajal" program, Spain is also acknowledged for partial financial support.

References and Notes

(1) Orengo, C.; Todd, A. E.; Thornton, J. M. *Curr. Op. Struct. Biol*. **1999**, *9*, 374−382.

(2) Thornton, J. M.; Todd, A. E.; Borkakoti, N.; Orengo, C. *Nature Struct. Biol. Struct. Genom. Suppl.* **2000**, 991−994.

(3) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. *Nucleic Acids Res.* **2000**, *28*, 235−242.

(4) PDB Statistics at http://betastaging.rcsb.org/pdb/statistics/holdings.do (accessed October 2007).

(5) Ding, C. H.; Dubchak, I. *Bioinformatics* **2001**, *17*, 349−358.

(6) Yu, C.-S.; Wang, J.; Yang, J.-M.; Lyu, P.-C.; Lin, C.-J.; Hwang, J.-K. *Proteins: Struct., Funct., Bioinf.* **2003**, *50*, 531−536.

(7) Prusis, P.; Muceniece, R.; Andersson, P.; Post, C.; Lundstedt, T.; Wikberg, J. E. *Biochim. Biophys. Acta* **2001**, *1544*, 300−357.

(8) Prusis, P.; Uhlen, S.; Petrovska, R.; Lapinsh, M.; Wikberg, J. E. S. *BMC Bioinformatics* **2006**, *7*, 167.

(9) A search performed using SciFinder Scholar with the phrases "more folded" or "less folded" produced 183 items. Most of the papers analyzed used the terms in a qualitative way, and only a tiny fraction uses measures like the radius of gyration to quantify the degree of folding.

(10) Randić, M.; Krilov, G. *Chem. Phys. Lett.* **1997**, *272*, 115−119.

(11) Randić, M.; Krilov, G. *Int. J. Quantum Chem.* **1999**, *75*, 1017−1026.

(12) Balaban, A. T.; Rucker, C. *J. Chem. Inf. Comput. Sci*. **2001**, *41*, 1145−1149.

(13) Liu, L.; Wang, T. *Int. J. Quantum Chem.* **2007**, *107*, 1970−1974.

(14) Estrada, E. *Chem. Phys. Lett.* **2000**, *319*, 713−718.

(15) Estrada, E. *Bioinformatics* **2002**, *18*, 697−704.

(16) Estrada, E. *Comput. Biol. Chem.* **2003**, *27*, 305−313.

(17) Estrada, E. *Proteins: Struct., Funct., Bioinf.* **2004**, *54*, 727−737.

(18) Estrada, E. *J. Chem. Inf. Comput. Sci* **2004**, *44*, 1238−1250.

(19) Estrada, E.; Uriarte, E. *Comput. Biol. Chem.* **2005**, *29*, 345−353.

(20) Estrada, E.; Uriarte, E.; Vilar, S. *J. Proteome Res.* **2006**, *5*, 105−111.

(21) Biedl, T.; Demaine, E.; Lazard, S.; Lubiw, A.; O'Rourke, J.; Overmars, M.; Robbins, S.; Streinu, I.; Toussaint, G.; Whitesides, S. *Discuss. Comput. Geom*. **2001**, *26*, 269−281.

(22) Quine, T.; Cross, T. A.; Chapman, M. S.; Bertram, R. *Bull. Math. Biol*. **2004**, *66*, 1705−1730.

(23) Cvetković, D.; Rowlinson, P.; Simić, S. *Eigenspaces of Graphs*; Cambridge University Press: New York, 1997.

(24) Lennard-Jones, J. E. *Proc. R. Soc. London* **1949**, *A198*, 1−13.

(25) Hall, G. G.; Lennard-Jones, J. E. *Proc. R. Soc. London* **1950**, *A202*, 155−165.

(26) Hall, G. G.; Lennard-Jones, J. E. *Proc. R. Soc. London* **1951**, *A205*, 357−374.

(27) Lennard-Jones, J. E.; Pople, J. *Proc. R. Soc. London* **1950**, *A202*, 166−180.

(28) Pauling, L. *Proc. Natl. Acad. Sci. U.S.A.* **1958**, *44*, 211−216.

(29) Trinajstić, N. *Chemical Graph Theory;* CRC Press: Boca Raton, FL, 1992; p 20.

(30) Estrada, E.; Guevara, N.; Gutman, I.; Rodríguez, L. *SAR & QSAR Environ. Res.* **1998**, *9*, 229−240.

(31) Estrada, E. *J. Chem. Inf. Comput. Sci*. **1999**, *39*, 90−95.

(32) Dix, D. B. An application of iterated line graphs to biomolecular conformation. http://www.math.sc.edu/∼dix/graph.pdf (accessed October 2007).

(33) Brown, R. D. *J. Chem. Soc.* **1953**, 2615−2621.

(34) We consider that the most folded conformation permits the largest overlap between dihedral orbitals, which reduces the dihedral energy to a minimum.

(35) Horn, R. A.; Johnson, C. R. *Matrix Analysis*; Cambridge University Press: New York, 1985.

(36) The current approach is not designed to find the optimal conformation of a protein chain by minimizing the dihedral energy along. However, this expression could be used in combination with other types of energy in such conformational calculations.

(37) Widom, B. *Statistical Mechanics. A Concise Introduction for Chemists*; Cambridge University Press: New York, 2002.

(38) Estrada, E.; Hatano, N. *Chem. Phys. Lett*. **2007**, *439*, 247−251.

(39) Abramowitz, M.; Stegun, I. A. *Handbook of Mathematical Functions with Formulas, Graphs and Tables*; National Bureau of Standards; Washington, DC, 1964; Applied Mathematics Series 55.

(40) Tilton, R. F.; Dewan, J. C.; Petsko, G. A. *Biochemistry* **1992**, *31*, 2469−2481.

(41) Arevalo, J. H.; Taussig, M. J.; Wilson, I. A. *Nature* **1993**, *365*, 859−863.

(42) Chen, J.; Wang, R.; Taussig, M.; Houk, K. N. *J. Org. Chem*. **2001**, *66*, 3021−3026.