A Protein Folding Degree Measure and Its Dependence on Crystal Packing, Protein Size, Secondary Structure, and Domain Structural Class

Ernesto Estrada*

Molecular Informatics, X-ray Unit, RIAIDT, Edificio CACTUS, University of Santiago de Compostela, 15782 Santiago de Compostela, Spain

Received November 28, 2003

Comparing two or more protein structures with respect to their degree of folding is common practice in structural biology despite the fact that there is no scale for a folding degree. Here we introduce a formal definition of a folding degree, capable of quantitative characterization. This enables ordering among protein chains based on their degree of folding. The folding degree of a data set of 152 representative nonhomologous proteins is then studied. We demonstrate that the variation in the folding degree seen for this data set is not due to crystallization artifacts or experimental conditions, such as resolution, refinement protocol, pH, or temperature. A good linear relationship is observed between the folding degree and the percentages of secondary structures in the protein. The folding degree is able to account for the small changes produced in the structure due to crystal packing and temperature. Automating the classification of proteins into their respective structural domain classes, namely mainly-α, mainly-β, and α-β, is also possible.

INTRODUCTION

One of the main goals in modern structural molecular biology is the finding of patterns and relationships between structures in genomes, transcriptomes, proteomes, and metabolomes.[1,2] The first key step in understanding the complex structures of any of these "omes" is in the characterization of their individual components. This characterization allows these components to be classified into groups, their similarities to be analyzed, their evolution understood, and their functions related, etc.[3,4]

The characterization of three-dimensional (3D) structures is of particular importance in the study of proteins.[5−7] The 3D structure of proteins is a complex physical[8] and mathematical[9,10] object, the understanding of which has been a major goal in modern molecular biophysics. Generally, interactions of a protein with other molecules are determined by those amino acids in close proximity regardless of their relatively large separation in sequence.[11] This implies that the majority of protein function information is given by the 3D structure rather than the sequence. In fact, many human diseases, such as age-associated degenerative diseases, are caused by "anomalies" in the 3D structure of some proteins that are abnormally folded.[12,13] Consequently, characterizing the folding of a protein is crucial in understanding its chemical and biological behavior.

In protein sciences, the term "folding" is used intuitively, i.e., it is used without a formal definition. Even more confusing is that it is used quantitatively despite the fact that no defined scale exists. For instance, a search for the exact match "more folded than" in papers related to proteins yields an extensive list of references covering physical, organic, and biological chemistry.[14−18] In some cases, experimental evidence is used to justify the comparison of folding between two or more structures, such as chemical shift dispersion in the ^{1}H NMR one-dimensional spectrum or high fluorescence maximum emission intensity. However, the main point here is that an expression such as "more than" is transitive, i.e., if A is more folded than B and B is more folded than C, then A must be more folded than C. This implies a ranking of folding such that each protein is characterized with a numerical value corresponding to its degree of folding. The term folding degree has also been used in chemical literature without definition.[18] This situation is also found in biological sciences where the comparison between folded structures is used for morphometric analysis of organs in intra- or interspecies studies.[19−21]

Recently we introduced a quantitative measure of folding for chains which has been useful in the study of proteins.[22−25] This index has not been formally related to the "concept" of the folding degree, as the latter has not been defined in the literature. Here we intend to fill this gap by defining the concept of the folding degree in a formal mathematical way. Introducing the concept of the folding degree is based on the intuitive idea of folding that is used in chemical and biological sciences. Using this formal definition, we are able to show that the index previously introduced is in fact a quantitative measure of the folding degree for (protein) chains. The folding degree index is then studied for a representative data set of 152 nonhomologous proteins. The influence of crystal packing, experimental conditions, such as resolution, refinement protocol, temperature, and pH as well as secondary structure on the folding degree of proteins are analyzed. This index is also used to classify proteins into their respective structural domains.

FOLDING-RELATED DEFINITIONS

We will start with an intuitive idea of folding. According to Oxford dictionary,[26] to fold is to bend (something) over on itself so that one part of it covers another. In the current

* Corresponding author fax: 0034-981-547 077; e-mail: estrada66@yahoo.com.

10.1021/ci034278x CCC: $27.50
Published on Web 04/21/2004

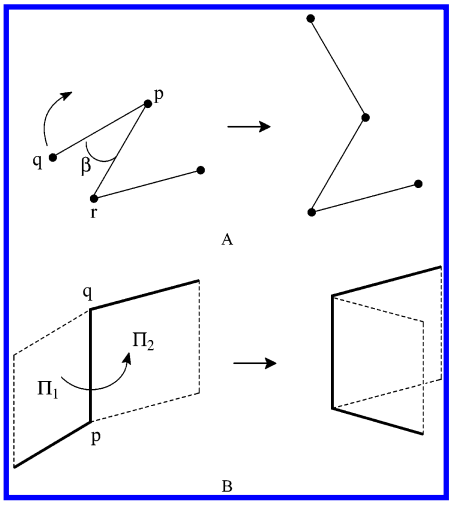

Figure 1. Examples of folding of a chain in the plane (2D folding) (A) and in the space (3D folding) (B).

work we are particularly interested in folding of polygonal chains. A polygonal chain, or simply a chain, is defined as follows.[27]

Definition 1. Let G = (V,E) be a connected graph having n vertices and $n - 1$ edges, i.e., it is a tree. Then if only two vertices in G have degree one, i.e., incident with only one edge, and the other $n - 2$ vertices have degree two, i.e., incident with two edges, the graph is a chain (also known as a path or a polygonal chain). We assume that all edges have fixed lengths.

Consider a chain C and choose an interior vertex p. Let q and r be the two vertices incident with p, as illustrated in Figure 1A. Thus the edges qp and qr form an angle β, defined as the amount of rotation about the vertex p required to bring one edge into correspondence with the other. A change of folding that consists on modifying the angles β for any of the interior vertices of the chain C on the plane will be called a 2D folding. In Figure 1A we give an illustration of this kind of folding. It can be formally defined as follows.

Definition 2.[9] A 2D folding conformation $Q = \{q_1, \cdots, q_n\}$ of the chain C is a continuous map ρ:C$\rightarrow\mathscr{R}^2$, i.e., a mapping of each vertex v_i to a point $q_i \in \mathscr{R}^2$, which *preserves all the lengths of the edges* of C. To avoid overcounting, ρ is considered up to any translation, rotation, and reflection of the plane.

From a chemical point of view a 2D folding change consists of changing bond angles in a molecule. It is well-known that bond lengths and bond angles only show very small variations in a molecule mainly due to vibrational motions. Consequently, to fold a molecule in the plane we need to drastically deform bond angles, changing dramatically hybridization of the corresponding atoms. This implies that **2D folding of a chain is chemically nonrealizable**.

As we are interested in folding which can be produced in real molecules, e.g., proteins, we will concentrate on chains in which bond lengths and bond angles are fixed. A good approximation to such chains is the *revolute chain*.[10] A revolute chain is a chain with fixed angles between pairs of incident edges in which only revolute motions around edges are permitted. A revolute motion is defined as follows.

Definition 3.[28] Consider a chain C with an interior edge pq. There are two subchains P and Q such as p $\in$ P and q $\in$ Q. If we keep P and Q individually rigid, and leaving P fixed in space, rotate Q around pq by an angle θ, maintaining the bond angles unchanged, we have produced a revolute motion at pq. We will refer to this motion as a **local revolute motion** around pq.

In Figure 1B we illustrate this kind of motion. As can be seen in this figure a revolute motion changes the angle between the planes formed by vertices spq and pqt. This angle is known as dihedral angle and it is defined as follows.

Definition 4.[29] The angle formed between two planes Π_1 and Π_2, which have normal vectors n_1 and n_2 is known as the dihedral angle between Π_1 and Π_2 and is given by the dot product of the normals: $\cos\theta = \hat{n}_1 \cdot \hat{n}_2$.

Consequently, a local revolute motion can also be named as a **local dihedral motion** about the corresponding edge. A 3D folding consists on modifying the dihedral angles θ for any of the interior edges of the chain C in the space through one or more local dihedral motions. A 3D folding conformation is then formally defined as:

Definition 5.[30] A 3D folding conformation $Q = \{q_1, \cdots, q_n\}$ of the chain C is a continuous map ρ:C$\rightarrow\mathscr{R}^3$, i.e., a mapping of each vertex v_i to a point $q_i \in \mathscr{R}^3$, satisfying the constraints that the angle between vectors $q_{i-1}q_i$ and q_iq_{i+1} is θ_i ($i \geq 2$), and the distance between q_i and q_{i+1} is d_i ($i \geq 1$). The points q_i and q_{i+1} are connected by a straight line e_i. To avoid overcounting, ρ is considered up to any translation, rotation and reflection of the space.

In this work we will consider C to be a protein chain with Ψ_i, ω_i, and ϕ_i being the torsional or dihedral angles of the backbone chain of the protein. Ψ_i describes the rotation about the $C_{\alpha i}$–C_i(=O) bond, and ω_i defines the rotation about the peptidic bond C_i(=O)–N_{i-1}. The normal trans planar peptide bond has $\omega_i = 180°$. ϕ_i describes rotation about the N_i–$C_{\alpha i}$ bond.

This implies that a chain C may be represented as a graph G = (V,E) , where $V = \{v_1, \cdots, v_n\}$ is the set of n vertices and E is the set of $n - 1$ edges in the graph. In this representation, atoms of the protein chain are vertices in G, and covalent bonds connecting these atoms are edges. This graph is specified by fixed bond angles θ_i at each vertex v_i, $i = 2, \cdots, n - 2$ and by fixed bond length d_i between vertices v_i and v_{i+1}, $i = 1, \cdots, n$.

FOLDING DEGREE

According to the definition of 3D folding, a conformation of C may be specified by the position of e_1 and dihedral angles θ_i, $i = 1, \cdots, n - 3$, where θ_i is the angle between planes $e_{i-1}e_i$ and e_ie_{i+1} ($i \geq 2$) which in the protein corresponds to Ψ_i, ω_i, and ϕ_i. Intuitively, a *folded conformation* is one where a part is bent over or doubled up such that it lies on another part.[26] Consequently, if we consider two planar folding conformations, Q and Q′, of a chain with four vertices $V = \{v_1,v_2,v_3,v_4\}$ the folded planar conformation will be that in which the plane v_1,v_2,v_3 "lies on" the plane v_2,v_3,v_4, i.e., the dihedral angle $\theta_1 = 0°$. We will call this conformation a *meander*. Thus the "less folded" conformation will be that where both planes lie apart from each other, i.e., when the dihedral angle $\theta_1 = 180°$. We will call this conformation a *zigzag*. This intuition leads to the definition of the *folding degree* for a revolute chain.

Definition 6. The folding degree is a continuous scale ranking the folding conformations of a chain C from a

maximal value corresponding to the hypothetically most folded conformation, i.e., for 3D folding the conformation for which $\theta_i = 0° \ \forall \ i$, to the less folded or zigzag conformation, i.e, for 3D folding the conformation for which $\theta_i = 180° \ \forall \ i$.

As we have seen the information needed to characterize the 3D folding degree is contained in dihedral angles. This information may be represented using a simple graph that only takes into account the dihedral angles θ_i, $i = 1, \cdots, n - 3$ of P. This graph will be denoted as the *third line graph* of G, $L^3(G)$,[31] because it can be obtained by the iterative line graph sequence procedure. Suppose G represents the chain C and $n \geq 4$. Let L(G) be the line graph of G, whose vertices are the edges of G and whose edges are unordered pairs of distinct edges of G which share a vertex. Let $L^2(G)$ denote the second line graph of G, i.e., the line graph of L(G), and $L^3(G)$ be the third line graph of G. The vertices of G represent the atoms of the chain, the vertices of L(G) represent pairs of adjacent atoms, i.e., bonds, the vertices of $L^2(G)$ represent triples of consecutive atoms, i.e., bond angles, and the vertices in $L^3(G)$ represent quadruples of consecutive atoms in the protein chain, i.e., Ψ_i, ω_i, and ϕ_i dihedral angles.

CRITERIA FOR MEASURING A FOLDING DEGREE

The most simple candidate for measuring a folding degree is a sum of dihedral angles in the chain. If we consider for instance a sum of cosines of dihedral angles we will obtain a maximum for a chain conformation in which all cosines of dihedral angles are equal to 1, i.e., $\theta_i = 0° \ \forall \ i$ or the hypothetical fully folded state. The minimum will be obtained if all cosines of dihedral angles are equal to -1, which corresponds to the fully extended conformation, $\theta_i = 180° \ \forall \ i$. To avoid size dependence we can define S_{DA} as the sum of cosines of dihedral angles in a chain of n vertices, i.e., $n - 3$ dihedral angles, divided by $n - 3$:

$$S_{DA} = \frac{1}{n-3} \sum_i \cos(\theta_i)$$

To test the appropriateness of this index as a measure of a folding degree we will consider the following chains. Let C and C′ be two chains with seven vertices embedded on a cube as illustrated in Figure 2. They will have fixed bond lengths, e.g., one as in a unit cube, and fixed bond angles, 90°. The sequence of dihedral angles in C is (90 0 90 0 90) and that in C′ is (0 90 90 90 0). It is clear that both chains are different and that they are not mirror images to each other. However, it is evident that $S_{DA}(C) = S_{DA}(C') = 2/5$. Thus, **the sum of cosines of dihedral angles in chains is not enough to differentiate the folding degree** of these chains. It is evident that there will be a larger number of such chains as the chain length increases, which makes this index useless to characterize the degree of folding.

The main difference between chains C and C′ is the "distribution" of dihedral angles along the chain. In chain C both meanders are only separated by a 90° dihedral angle at the center of the chain. On chain C′ both meanders are on the extremes of the chain and are separated by three 90° dihedral angles. This larger separation between both meanders in C′ with respect to C makes that the "unfolded" region,

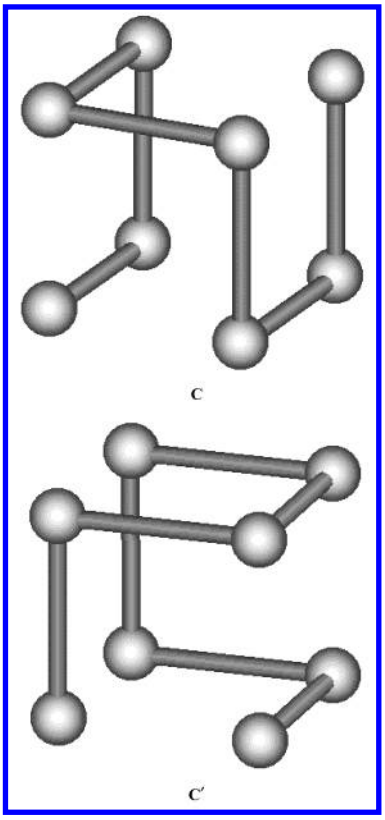

Figure 2. Two chains (with fixed bond lengths and bond angles) of seven vertices having the same sum of dihedral angles but having a different degree of folding.

i.e., that consisting of 90° dihedral angles, is longer in C′ than in C. In consequence, we can consider that C′ is less folded than C on the basis that it contains a longer "unfolded" region. In closing, we have to say that according to the criteria we are following in this work a folding degree measure has to take into account not only the number of dihedral angles but also their distribution along the chain.

If we consider the case of proteins we will find that a folding degree criterion commonly used is the percentage of helix and strand in the protein. Helical parts of a protein are more folded than regions forming strands. Consequently, if protein P has a higher percentage of helix than another, say P′, then P can be considered to be more folded than P′. However, these percentages of helix and strand say nothing about the distribution of these helices or strands along the protein chain. On the other hand, these percentages do not account for the differences among types of helices, such as α, 3_{10}, or π, or for the differences in folding of other secondary structures, such as turns. It is possible to extend the percentages of helix and strand to account for all these other structures, but even in this case this larger number of "descriptors" do not account for the distributions of such secondary structure elements along the protein chain.

A situation similar to that previously described is found for several other molecular attributes. If we take, for instance, the degree of branching for alkanes such as *n*-pentane (5), isopentane (2M4), and neopentane (22MM3) we intuitively recognize that the branching increases in the order 5 < 2M4 < 22MM3. We can express this order in terms of the percentage of methyl groups present in the structures: 5 (40%), 2M4 (60%), and 22MM3 (80%). The same criterion can be extended to the percentages of other groups, e.g.,

$\%CH_2$, %CH, and %C. However, these criteria do not distinguish between certain set of structures, such as 2-methylpentane and 3-methylpentane, which have the same values for $\%CH_3$ (50%), $\%CH_2$ (33.3%), %CH (16.7%), and %C (0%). These compounds have different properties, which are expected to depend on branching degree, which makes this measurement based on percentages useless for the purpose of characterizing branching degree. The point here as well as in the case of the folding degree is that *not only the number of local attributes (branching points or folded regions) in the molecule but also its distribution across the molecule can be important for characterizing the global molecular attribute (the branching or folding degree)*. As a consequence we need more complex quantitative measures to characterize the branching or folding degree of (protein) molecules.

A GRAPH SPECTRAL MEASURE OF FOLDING DEGREE

According to the previous analysis we need a folding degree measure that accounts not only for the sum of dihedral angles but also for their distribution along the chain. The general form of this measure could be as follows

$$M = a_0 \sum_i f_i + a_1 \sum_{ij} f_{ij} + a_2 \sum_{i,j,k} f_{ijk} + \cdots \qquad (1)$$

where f_i is a function of the dihedral angles, e.g., $\cos(\theta_i)$, f_{ij} is a function of pairs of contiguous dihedral angles, f_{ijk} is a function of triples of contiguous dihedral angles and so forth, and a_i are coefficients accounting for the "importance" of each term. It is straightforward to realize that if we take $f_i = \cos(\theta_i)$ and $a_0 = 1$, the first term of the sum is S_{DA}. Then, the rest of the terms can be defined to account for the distribution of the dihedral angles along the chain in a decreasing way. A simple way of doing so is to take $a_k = 1/k!$, which also simplifies the mathematical procedures to be used (see further). Returning to the graph theoretical approach the function f_i can be considered as a weight assigned to the vertex i of the graph, i.e., a "vertex weight".

We recall that we are considering here the third line graph of the molecular graph $L^3(G)$, in which each vertex is weighted by a function f_i of dihedral angles. The "weighted" adjacency matrix of $L^3(G)$ is defined as follows.

Definition 7. Let **C** be a diagonal square matrix whose diagonal elements are $c_{ii} = f_i$ and nondiagonal elements are zeroes. Let $\mathbf{A} = \mathbf{A}[L^3(G)]$ be the adjacency matrix of the third line graph of the molecular graph representing the backbone chain of a protein. Then, the weighted adjacency matrix of $L^3(G)$ is defined as $\mathbf{M} = \mathbf{A} + \mathbf{C}$.

A mathematical operator that accounts for the contribution of vertex weights along a chain is the spectral moment of the adjacency matrix associated to the graph. The spectral moment of order k is defined as the trace, **Tr**, of the kth power of M

$$\mu_k = \mathbf{Tr}(M^k) \qquad (2)$$

where the trace is the sum of the diagonal elements of the corresponding matrix.

It has been previously shown that spectral moments of a matrix accounts for contributions of the different (weighted) fragments in the graph.[32−34] For instance, $\mu_1 = \sum_i f_i$ accounts for local contributions centered on vertices, $\mu_3 = \sum_i f_i^3 + 3\sum_r (f_i + f_j)_r$ take into account contributions coming from the r pairs of adjacent vertices, i.e., the term $\sum_r (f_i + f_j)_r$ in the expression of μ_3 can be considered as a component of the term f_{ij} in (1). Higher order spectral moments accounts for contributions coming from sequences of consecutive weighted vertices. Thus, we can use spectral moments to account for the distribution of weights along a chain as given by expression 1, which permits to define a folding degree index.

Definition 8. Let P be a (protein) chain of n vertices and let **M** represents the weighted adjacency matrix of $L^3(P)$. Then a folding degree index is defined as follows:

$$I_3 = \frac{1}{n-3} \sum_{k=0}^{\infty} a_k \mu_k = \sum_{k=0}^{\infty} \frac{\mu_k}{k!} = \sum_{k=1}^{\infty} \frac{\mu_k}{k!} \qquad (3)$$

Note that the normalization factor $1/n-3$ has been taken into account because if we carry out the sum from $k = 0$ we are counting the number of vertices in $L^3(P)$ due to $\mu_0 = n - 3$. In two previous works concerning this folding degree index,[23,24] by mistake, we used in the expressions for I_3 a normalization factor of $1/n$ instead of $1/n-3$. However, all the calculations were carried out using an expression identical to (3), that is using the normalization factor of $1/n-3$. Other descriptors that characterize 3D sequence of proteins have been proposed in the literature, which have been valuable for studying lattice proteins.[35−37]

The subscript 3 in I_3 is used to realize that this folding degree index is referred to 3D folding as defined in this work. An I_2 index can also be defined in a similar way but using bond angles instead of using dihedral angles in order to account for the 2D folding degree, which however has been found to be of low importance from a chemical point of view (nonpublished results).

To avoid truncating the infinite summation in this expression we take advantage of the relationship between spectral moments and eigenvalues of a matrix and transform (3) into the following expression:[22,23]

$$I_3 = \frac{1}{n-3} \sum_{k=0}^{\infty} \left(\frac{\sum_{i=1}^{n-3} (\lambda_i)^k}{k!} \right) = \frac{1}{n-3} \sum_{i=1}^{n-3} e^{\lambda_i} \qquad (4)$$

Using expression 4 the folding degree is calculated on the basis of the eigenvalues of adjacency matrix of the third line graph weighted by cosines of dihedral angles, **M**. The matrix **M** is a tridiagonal symmetric matrix which reduces dramatically the complexity for the eigenvalues calculation compared with other adjacency matrices of the same order. For instance, the QL algorithm, which has the complexity of $O(n^3)$ per iteration for a general matrix, is only $O(n)$ per iteration for a tridiagonal matrix.[38] On the other hand, Gu and Eisenstat have proved that all eigenvalues of a tridiagonal matrix can be found in $O(n \log_2 n)$ operations using a divide-and-conquer algorithm.[39]

AMINO ACIDS CONTRIBUTIONS

According to the proper definition of spectral moments the term $\sum_{r=0}^{\infty} (\mu_r/r!)$ in expression 3 can be written as a sum

of the form given below[25]

$$\sum_{r=0}^{\infty}\frac{\mu_r}{r!}=\sum_{r=0}^{\infty}\frac{\mu_r^{(1)}}{r!}+\sum_{r=0}^{\infty}\frac{\mu_r^{(2)}}{r!}+\sum_{r=0}^{\infty}\frac{\mu_r^{(31)}}{r!}+\cdots+\sum_{r=0}^{\infty}\frac{\mu_r^{(n-3)}}{r!} \quad (5)$$

where $\sum_{r=0}^{\infty}(\mu_r^{(i)}/r!)$ is the contribution of vertex i in $L^3(G)$ to our global folding degree index.

The terminal amino acid contributes to I_3 with only one dihedral angle, Ψ_i, while the rest of the amino acids contribute with two dihedral angles, Ψ_i and ϕ_i. Dihedral angles corresponding to peptide bonds, ω_i, are taken to be noncontributing to any of the amino acids which are connected by them.

Therefore, the contribution of any amino acid in the protein to the folding degree index I_3 can be obtained using the following expressions[25]

$$R_i=\sum_r\frac{\mu_r^{(3i-3)}}{r!}+\sum_r\frac{\mu_r^{(3i-2)}}{r!} \quad (6)$$

$$R_1=\sum_r\frac{\mu_r^{(1)}}{r!} \quad \text{and} \quad R_{n-3}=\sum_r\frac{\mu_r^{(n-3)}}{r!} \quad (7)$$

for nonterminal and terminal amino acids, respectively. The peptide bonds will have contributions determined by the following expression:[25]

$$R_{i-j}=\sum_r\frac{\mu_r^{(2i+j-2)}}{r!} \quad (8)$$

However, due to the fact that most peptide bonds have a value of the dihedral angle approximately equal to 180° (trans conformation), the values of R_{i-j} will be almost constant for all peptide bonds (the only one exception is that of *cis*-proline).

WORKING DATA SET

We studied our folding degree index of a representative data set of 152 nonhomologous, high-resolution protein structures as determined by X-ray crystallography. These proteins ranged between 50 and 753 amino acid residues.[40] The 3D structures of these proteins were solved at less than 2.0 Å of resolution and an R factor of ≤ 0.2. The PDB entries for these proteins are as follows: 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, 2cba, 1arb, 2dri, 1s01, 2ebn, 1tml, 1nar, 1amp, 1fnc, 2glt, 8abp, 2ctc, 1gca, 1sbp, 1ede, 1hvd, 1ads, 1trb, 8tln, 1tca, 1frp, 1pbp, 2pia, 1apv, 1bmd, 1smr, 1apm, 1hle, 1gox, 2bbk, 2mnr, 1fba, 2ohx, 2sil, 1pbe, 1php, 1oyb, 1chm, 4enl, 2cts, 1alk, 1pii, 2hpd, 2pgd, 3cox, 2olb, 1ddt, 1aoz, 1gof, 1trk, 1cdg, 8acn.

FOLDING DEGREE: PACKING, COMPACTNESS, OR SOMETHING DIFFERENT?

To characterize molecular attributes is not an easy task. The first problem we can find is that most of these attributes are not well defined. In the case of shape attributes the situation is not different. We can find, for instance, references to "compactness", "packing", or "folding" without making any differentiation among them. In fact, one of the referees of this work has recognized that "*people understand folding to be related to how globally compacted into say a ball-shaped region of space a protein might be, with the extend of folding greater the smaller the ball and the greater the length of the polymer*". This definition appears to be, however, more related to the concept of compactness than with that of folding.

Our objective in this section is to analyze whether the information contained in the folding degree index I_3 is a duplication of information already contained in other shape attributes, such as packing and compactness. In doing so, our approach is simple and straightforward. We will correlate our folding degree index with those describing packing and compactness. The lack of correlation, say a correlation coefficient below 0.50, will be understood as a statistical independence between the corresponding indices and consequently as a lack of duplicated information in such descriptors.

Packing[41] is understood as the placement of objects so that they touch in some specific manner, often inside a container with specified properties. On the other hand, and in a very similar fashion **compactness**[26] is defined as the spatial property of being crowded together. Something is compact when it is closely and firmly united or packed together, dense, or occupying little space compared with others of its type.

Three packing/compactness descriptors that will be analyzed here are the following. Liang and Dill defined total protein packing, P_t, as the ratio of the van der Waals volume divided by the sum of the envelope volume and the pocket volume.[42] Occluded surface packing, OSP, defined by Pattabiraman, Ward, and Fleming measures the interatomic occluded surface area for each atom in the protein.[43] Finally, the radius of gyration of the protein, R_G, has been widely used as a measure of the global compactness and in some cases also as a measure of packing.[44]

In the first case we plot our folding degree index, I_3, versus P_t values reported by Liang and Dill for 15 proteins ranging from 41 to 346 residues.[42] In the case of the other two parameters we plot the values of I_3 versus OSP and R_G for 152 proteins studied in the current work. We illustrate the three plots in Figure 3, where it is evident the lack of relation between packing/compactness indices and the folding degree index I_3 ($R<0.2$). The same lack of correlation has been observed when we plotted P_t, OSP, and R_G versus the percentages of helix and strand for the corresponding proteins. These results clearly point to the conclusion that the folding degree index defined here appears to be independent from molecular descriptors accounting for packing and compactness. More importantly, we think that these

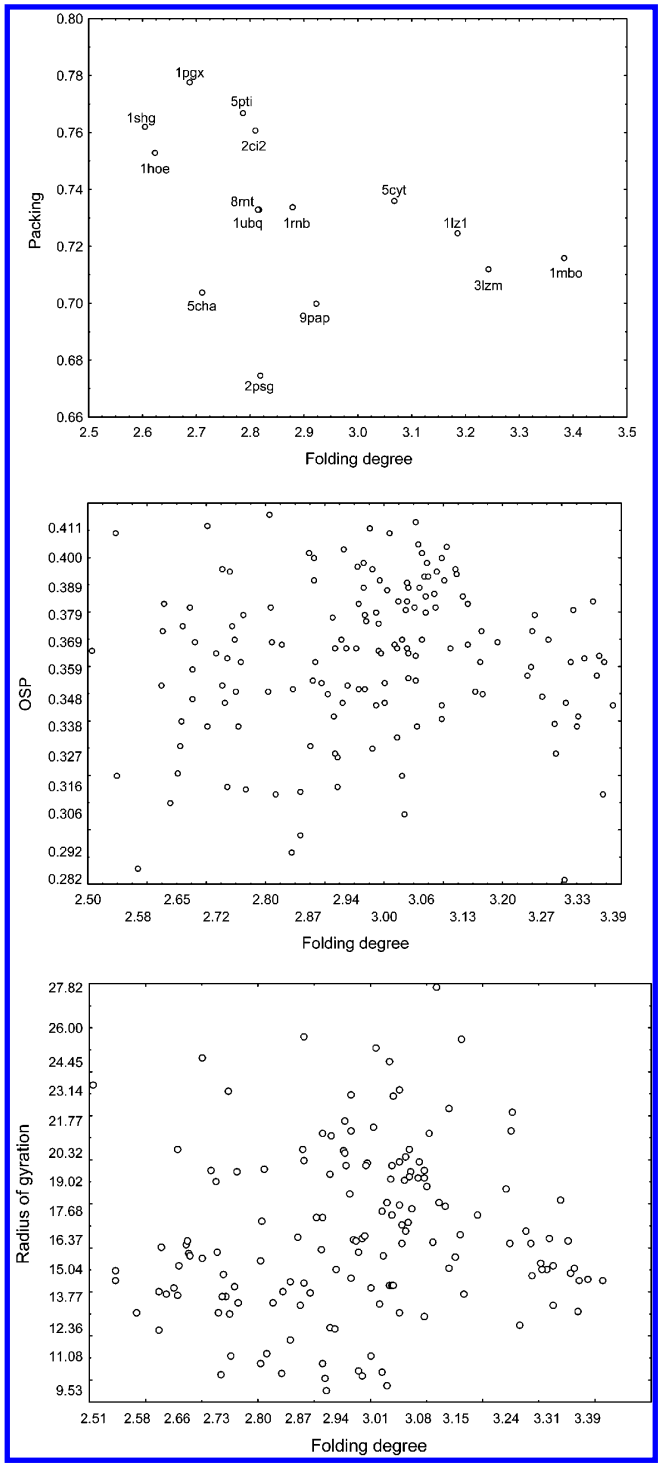

Figure 3. Relationships between the folding degree of real proteins (measured by I_3 index) and packing (P_t), occluded surface packing (OSP), and compactness measured by the radius of gyration (R_G).

results can help to understand that the folding degree is an independent shape attribute not related to the concepts of packing and compactness.

I_3 AND CRYSTAL PACKING

The first objective is to determine whether the variation of our folding degree index for the nonhomologous proteins studied is due to the intrinsic nature of the folding degree of these proteins or to effects produced by artifacts of crystallization. Our folding degree index varies as much as 27% for proteins in this data set with a standard deviation of 0.2063. The second objective is to observe whether our folding degree index is able to correctly account for changes produced by crystal packing. It is known that crystal packing may lead to a change in the number of contacts between neighboring molecules in the crystal, which in turn can influence the folding degree of the protein.[45−48] These contacts have no evolutionary role and therefore are significantly different from protein−protein interactions of a biological nature.[49]

Table 1. Folding Degree of Hen Lysozyme with Different Crystal Groups

PDB	group	resol	I_3	PDB	group	resol	I_3
4lyt(A)	$P2_1$	1.9	3.2080	1lza	$P4_32_12$	1.6	3.2364
4lyt(B)	$P2_1$	1.9	3.2119	193l	$P4_32_12$	1.33	3.2361
5lym(A)	$P2_1$	1.8	3.2250	194l	$P4_32_12$	1.4	3.2407
5lym(B)	$P2_1$	1.8	3.2167	1hel	$P4_32_12$	1.7	3.2534
2lzt	$P1$	1.97	3.2098	6lyt	$P4_32_12$	1.9	3.2366
1aki	$P2_12_12_1$	1.5	3.2346	2lym	$P4_32_12$	2.0	3.2491
				1lse	$P4_32_12$	1.7	3.2496

Table 2. Folding Degree of Pairs of Proteins with Different Crystal Groups

PDB codes	space groups	I_3
1une/1mkt	$P2_12_12_1/P3_121$	3.2196/3.1925
1pga/1pgb	$P2_12_12_1/P3_121$	2.7411/2.7350
1mku/1mks	$P2_12_12_1/P3_121$	3.2838/3.1812
1svn/1jea	$P2_1/P2_12_12_1$	2.9746/2.9797
1aaj/2rac	$P2_1/P1211$	2.6226/2.5929
1mpb/1mpc	$P2_1/P1$	3.1000/3.1000
1cub/1cuc	$P12_11/P2_12_12_1$	3.1430/3.1400

First we will investigate to what extent crystal packing is responsible for the observed 27% variation in the folding degree index. In doing so, we have selected hen lysozyme structure in four different crystal groups. As we can see in Table 1 the values of the folding degree vary between 3.2080 and 3.2534. This is within the range of variation for our folding degree index of lysozymes, e.g., 3.2080 for 4lyt(A) to 3.2540 for 1hhl. The percentage of variation due to differences in crystal packing for lysozyme is only 1.4%, significantly different from the variation observed for the whole data set of proteins studied here. Similar small changes in the folding degree are also observed for several other pairs of proteins in different crystal groups. The largest variation is observed for pair 1mku/1mks with crystal groups $P2_12_12_1$ and $P3_121$, respectively (see Table 2). However, this variation is only 3.1%. In closing, the variation in the folding degree that might be introduced by crystal packing is unable to account for the variation in the folding degree observed in the whole data set of nonhomologous proteins.

The number of crystal contacts is known to be small when compared with biological contacts. Most of the proteins solved by X-ray analysis and deposited in the PDB have three or more crystal contacts, but some have more than 20. Interfaces of these contacts are also relatively small. According to Carugo and Argos[46] 45% of these contacts have <100 Å^2 of interface and only 8% have contacts larger than 500 Å^2. These contacts are capable of modifying the folding degree significantly. For instance, the average folding degree for hen lysozyme belonging to crystal group $P2_1$ (see Table 1) is 3.2154 ($\pm$ 0.0073) and in group $P4_32_12$ it is 3.2431 ($\pm$ 0.0074). According to the *t*-test, the means of these two

groups are significantly different with a p-level of 0.0002, that is the probability of error involved in accepting that the two groups are significantly different in their folding degree. In other words, the folding degrees of hen lysozyme in monoclinic and tetragonal crystal groups are significantly different with a probability of 99.9998%. Considering 95% of confidence monoclinic structures have folding degrees in the range 3.2037−3.2270 and tetragonal structures in the range 3.2363−3.2499.

To understand whether crystal packing differences are responsible for the variation in the folding degree or whether this variation is responsive to experimental errors, we investigate the change of the folding degree of proteins in the same crystal group. For instance, three structures of lysozyme from *E. coli* infected with bacteriophague T4 in crystal group $P3_221$ at 1.7 Å resolution show the following I_3 values: 3.2429 (2lzm), 3.2417 (3lzm), and 3.2409 (4lzm), giving a standard deviation of 0.0010. Similarly, small standard deviations are obtained for other proteins in the same crystal group, examples include rubredoxin from *P. furiosus* in $P2_12_12_1$ which has a standard deviation of 0.0047: 2.9140 (1brf), 2.9075 (1bq8), 2.9048 (1caa); pseudoazurin from *A. faecalis* in *P*65 which has a standard deviation of 0.0006: 2.7300 (1paz), 2.7293 (3paz), 2.7288 (1pza); subtilisin from *B. amyloliquefociens* in *P*21 which has a standard deviation of 0.0072: 2.9789 (1sbh), 2.9711 (1yja), 2.9854 (1yjb); and concavalin A from *C. ensiformis* in *I*222 which has a standard deviation of 0.0039: 2.6342 (1scs), 2.6375 (2ctv), 2.6298 (1enr). As can be seen from these values, the differences arising from experimental conditions (see further section for more extended analysis) are not responsible for the variation observed in the values of hen lysozyme in crystal groups $P2_1$ and $P4_32_12$. The significant differences between these two groups of crystal forms could be due to the different number of contacts produced by the packing of the proteins in such environments. For instance, the number of residues involved in crystal contacts in 1hel, lysozyme in the tetragonal group, is 33, while this number is only 4 for the monoclinic structure, 4lyt, in correspondence with the values obtained for our folding degree index. Similar variations are observed in Table 2 for three pairs of structures containing one orthorhombic ($P2_12_12_1$) and one trigonal ($P3_1$-21) component. In all these three cases, proteins in orthorhombic crystals show higher folding degrees than those in trigonal ones. However, there are some other pairs in different crystal groups that show no difference in the folding degree, such as those in $P2_1$ and *P*1.

It has been observed that protein−protein interactions affect secondary structure elements of proteins differently. Jones and Thornton[50] have observed that 53% of the interface residues in protein−protein interactions were classed as α-helix and 22 as β-sheet. Here we have observed that the majority of modifications in protein secondary structure produced by crystal packing occur in helices and turns of the proteins. β-sheets are almost unaffected by such an effect. For instance, Figure 4A plots the differences in amino acid contributions of two mainly α proteins, 1mku and 1mks, with crystal groups $P2_12_12_1$ and $P3_121$, respectively. Most of the changes in the folding degree observed are produced in the central turn as well as in some helices, but the β-sheets are unaltered. Similarly, the same plot for two mainly-β proteins (Figure 4B), 1pga and 1pgb, show marked changes occurring

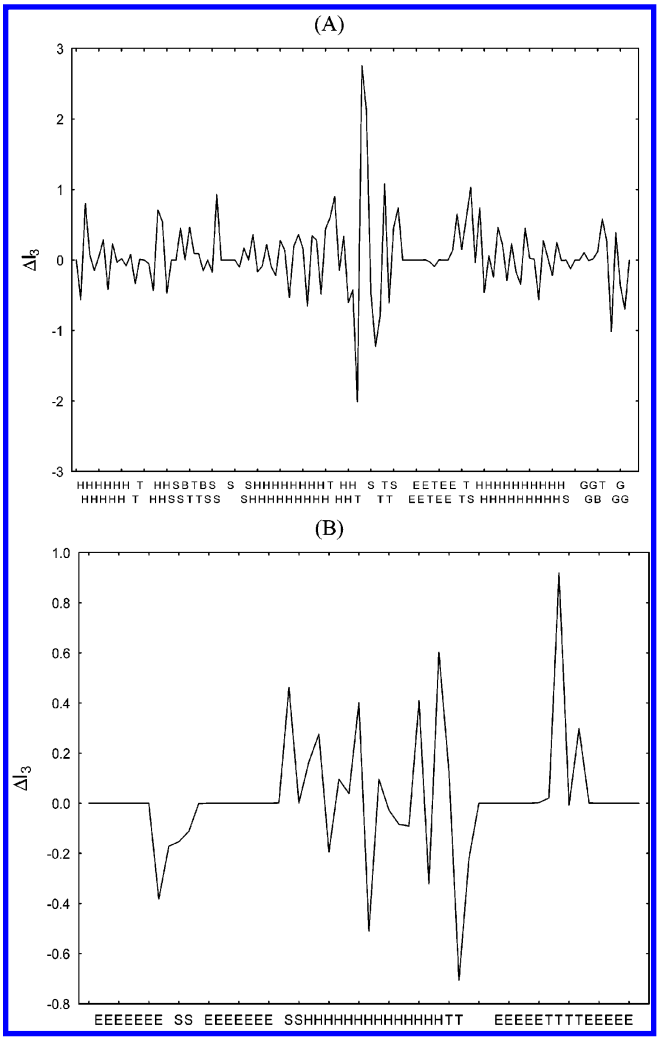

Figure 4. (A) Difference of amino acid contributions to the folding degree of two crystal forms of a mainly-α protein (phospholipase A2) in orthorhombic (1mku) and trigonal (1mks) forms. (B) Difference of amino acid contributions to the folding degree of two crystal forms of a mainly-β protein (protein G) in orthorhombic (1pga) and trigonal (1pgb) forms.

in both the central helix and the four turns, yet none of the β-sheets are altered by crystal packing.

The effect of crystal packing on the folding degree requires further investigation to obtain substantiated conclusions on this topic. Two conclusions can be derived from the current one. The first is namely that the variation observed in the folding degree of the data set studied is not accounted for by the variations produced by crystal packing. The latter effect accounts only for about 3% of the variation of the folding degree compared with the 27% change in the whole data set. The second conclusion is that in some cases, the effect produced by crystal packing on the molecular structure of proteins does affect the folding degree in a manner that is statistically meaningful.

I_3 AND EXPERIMENTAL CONDITIONS

Although experimental conditions are not directly related to the protein structure, they are able to modify the structure to an extent. The experimental conditions we will study here include the influence of resolution, refinement protocol, pH, and temperature on the folding degree of proteins. As previously, the analysis of the influence of these factors on

Table 3. Temperature Dependence of the Folding Degree and Other Macromolecular Descriptors, Such as Occluded Surface Packing (OSC) and Radius of Gyration (R_G)

PDB code	T (K)	OSP	R_G	I_3
1rat	98	0.3669	14.0843	2.8496
2rat	130	0.3589	14.1181	2.8485
3rat	160	0.3643	14.1068	2.8452
4rat	180	0.3608	14.1387	2.8480
5rat	220	0.3590	14.1240	2.8420
6rat	240	0.3623	14.1184	2.8353
7rat	260	0.3582	14.1392	2.8337
8rat	300	0.3525	14.2219	2.8384
9rat	320	0.3534	14.1746	2.8326
SD		0.0047	0.0406	0.0067
R		0.826	0.815	0.905

the values of the folding degree is 2-fold. On one hand, we want to know what influence experimental conditions have on the variability of the folding degree observed in the whole data set. Second we want to establish whether the folding degree is able to correctly explain some of the variations produced by these factors.

The first experimental conditions to explore are the resolution and refinement protocol. Ribonuclease A from *B. taurus* solved at seven different resolutions from 1.4 to 2.0 Å and by four different refinement protocols, RESTRAIN,[51] TNT,[52] X-PLOR,[53] and PROFFT, is used as an example.[54] The standard deviation for these seven structures is 0.0072 (1rpg: 2.998; 3rn3: 2.992; 1bel: 2.995; 1afk(A): 3.001; 1xps(A): 3.002; 1rnx: 2.982; 1aqp: 3.002), which is in the same range than that obtained when proteins in the same crystal group but solved by different laboratories were studied in the previous section. The variation produced by these different experimental conditions is only 0.3%, well below the percentage of variation obtained for the whole data set (27%).

Another important factor that may influence the folding degree of proteins is pH at which their structures are solved. To investigate the effect of this factor on the folding degree of proteins we selected structures of ribonuclease A studied at six different pH: 5.2 (1kf2), 5.9 (1kf3), 6.3 (1kf4), 7.1 (1kf5), 8.0 (1kf7), 8.8 (1kf8).[55] The standard deviation of the folding degree for these proteins is only 0.0005, and the percentage of variation is 0.04%: I_3 values are 2.8313, 2.8309, 2.8305, 2.8301, 2.8313, 2.8302. Clearly this factor does not make a significant contribution to the variation observed in the whole data set of nonhomologous proteins.

The last factor we are going to investigate in relation to the folding degree is temperature. It is anticipated that increasing the temperature will produce a defolding of the protein decreasing the folding degree. Ribonuclease A has been studied at nine different temperatures ranging from 98 to 320 K.[56] The folding degree values for these proteins are reported in Table 3 where the standard deviation produced by this factor is shown to be 0.0067, which again is in the same interval as obtained before. The percentage of variation introduced by these factors in nine proteins is only 0.6%. As expected, our folding degree index decreases when temperature increases and a straight line regression is obtained with a correlation coefficient of 0.91. For comparison, the correlation coefficients obtained by the changes in OSP[40,43] and the radius of gyration[44] with the increase of temperature are also shown. As can be seen our folding degree index best describes changes in the three-dimensional structure of ribonuclease A at different temperatures.

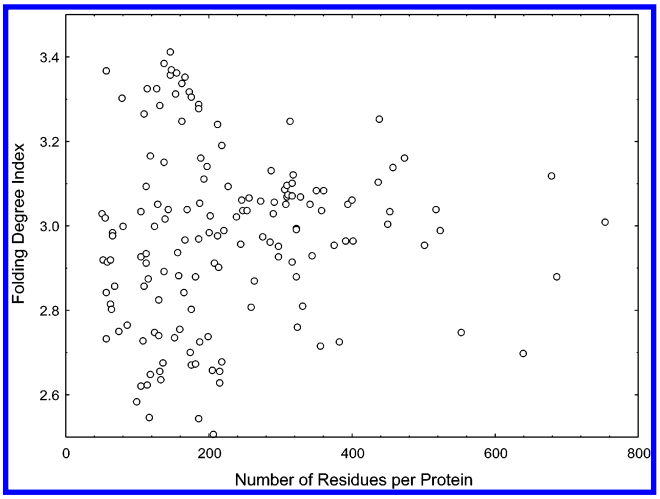

Figure 5. The protein folding degree for each of the 152 nonhomologous proteins as a function of the total number of residues per protein chain.

In closing, it is clear that the observed variation of the folding degree for the 152 proteins studied here is not due to crystallization artifacts or to experimental conditions but mainly due to the intrinsic nature of folding of these proteins.

I_3 AND PROTEIN SIZE

Figure 5 illustrates the plot of our folding degree index versus the number of residues per protein for the 152 nonhomologous proteins studied. As seen in this figure, there is no dependence between both magnitudes, which is interpreted as the protein size independence of our folding degree. This could be a consequence of the way in which we have defined the I_3 index as we have normalized the index dividing it by the number of residues. As a result, this index represents the average folding degree per residue in the protein. However, other average indices measuring the shape of a protein also show a marked dependence with the protein size. For instance, it has been shown that the average occluded surface packing (OSP)[40,43] increases markedly with the number of residues per protein up to approximately 200 residues and then increases only slightly with larger proteins for the same data set studied here. We wanted to explore this type of relationship further and found that the radius of gyration shows linear dependence with a number of residues per protein. It shows a dependence like that of the OSP but is less marked (graphics not shown).

As seen in Figure 5, small proteins have more variability in our folding degree than larger ones. In fact, the 82 proteins with less than 200 residues per protein have folding degrees ranging from 2.5 to 3.4, while those having more than 200 residues have folding degrees predominantly in the range of 2.7 and 3.2. Despite both groups having average I_3 values which are very close together (not statistically different according to *t*-test), the first has a standard deviation of 0.2443 while the second has a standard deviation of 0.1536, thus the greatest variability in the folding degree occurs in smaller proteins. It is worth noting that the standard deviation for the whole data set is 0.2063.

This observation can be rationalized in a similar way as that for OSP,[40] i.e., on the basis of the observation made by

Table 4. Folding Degree of Different Domains in Multidomains Proteins

residues	region	domain	I_3
	2gst		
217	1−217		3.1906
85	1−85	alpha-beta	2.9681
105	86−190	mainly alpha	3.3473
27	191−217	alpha-beta	3.0077
	8tln		
316	1−316		3.0712
149	6−154	alpha-beta	2.9302
162	155−316	mainly alpha	3.1963
	1ddt		
523	1−523		2.9790
187	1−187	alpha-beta	2.9523
180	200−379	mainly alpha	3.2637
144	380−523	mainly beta	2.5700

Privalov[57] that the maximum size of the protein domains approaches 200 residues. As a consequence larger proteins tend to be combinations of such domains. In the data set studied here, 57% of the proteins containing more than 200 residues have more than one domain. This percentage rises to 64% when the 7995 PDB entries reported in 1999 were analyzed by Dengler et al.[58] They studied a total of 13 767 chains and 18 896 domains. In contrast, only 6% of the proteins studied here with less than 200 residues are multidomain. In multidomain proteins the domains are linked together through the so-called *linkers* (http://mathbio.nimr.mrc.ac.uk),[59] which are flexible peptide chains of a few residues. In some cases the multidomain proteins show a folding degree which is a sort of average of the folding degrees of the different domains contained in them. Thus, extreme values normally found in mainly alpha or mainly beta domains are not present in these proteins. The linkers can also play a role in "averaging" the folding degree of the different domains due to their nonrigid character. Table 4 illustrates this averaging effect for three different multidomain proteins of more than 200 residues. If we analyze the third protein (PDB: 1ddt), for example, we can see that it contains regions with "extreme" values of the folding degree corresponding to the mainly α and mainly β domains. However, the total folding degree of the protein is an average value, which is in the range characteristic of proteins with more than 200 residues (see Figure 5), i.e., between 2.7 and 3.2.

The types of domains forming large proteins may potentially account for the difference in variability observed in the folding degree of proteins as a function of their size. If we consider one-domain proteins with more than 200 residues, we can see that 81% of them consist of $\alpha-\beta$ domains. This percentage is only 23% among the one-domain proteins with less than 200 residues. Surprisingly, despite the longer chains which have more alternative folds available to them, they are formed in a way in which the folding degree is bounded by narrow margins showing less variability than small proteins. From Table 4, a relationship between the folding degree index and the class domains in proteins can be inferred. As class domain classification is mainly based on the secondary structure of proteins, we will investigate the relationship between it and our folding degree index of the 152 nonhomologous proteins studied.

I_3 AND SECONDARY STRUCTURE

Secondary structure of a protein is the path of the polypeptide main chain through the three-dimensional space. It includes the formation of standard conformations, such as helices and sheets, by hydrogen bonds only between the atoms of the backbone. Consequently, the secondary structure of a protein is directly related to the degree of folding of the main chain. A semiquantitative characterization of the secondary structure is normally carried out through the use of the percentages of helix and percentage of strands in the protein. A protein is considered more folded than another if it contains a greater percentage of helix and/or a lesser percentage of strand. This criteria does not introduce a definitive order of the folding degree. For example, the following order of the folding degree is produced from the percentage of helix: 1cus (44%) < 1tph (45%) < 2ak3 (46%), yet the reverse order is obtained with the percentage of strand: 2ak3 (17%) < 1tph (16%) < 1cus (15%). This criterion is based on the consideration that β sheets are the least folded and helices are the most folded structures in proteins. Sheets show similarity with the fully extended chain ($\theta_i = 180° \; \forall \; i$), which is the least folded conformation of a revolute chain P. On the other hand, helices are intuitively considered as the most folded structures in proteins due to their dissimilarity with the hypothetical fully extended chain, which is analogous to say that they are more similar to a hypothetical chain for which $\theta_i = 0° \; \forall \; i$. It is easy to realize that this hypothetical chain is impossible to build due to the superposition of some vertices in the chain.

The values closest to 0° for the dihedral angles Ψ and ϕ observed in proteins are those corresponding to the different types of helices. Consequently, the most folded conformations found in a peptidic chain are those for 3_{10} helices ($\Psi = -49°$ and $\phi = -26°$), right-handed α helices ($\Psi = -57°$ and $\phi = -47°$), and π helices ($\Psi = -57°$ and $\phi = -70°$),[60] which indeed show the highest folding degree.[23] On the other hand, the most extended conformations, parallel and antiparallel β sheets, are the least folded structures in a protein, with dihedral angles close to 180°.

Figure 6 illustrates our folding degree index versus the percentages of helix and the percentage of strand for the 152 nonhomologous proteins studied. Our folding degree index shows a good linear correlation with both the percentage of helix ($R = 0.928$) and the percentage of strand ($R = -0.918$). Other macromolecular descriptors, such as OSP[40,43] or the radius of gyration, show no correlation with the percentage of secondary structure (correlation coefficients lower than 0.2 in all cases as shown before).

As seen in Figure 6, the folding degree index increases with the increase in the percentage of helix in the protein and decreases with a lower percentage of strand. This makes the intuitive idea of secondary structure content compatible with the quantitative and exact measure of the folding degree. According to the square correlation coefficients the I_3 index contains at least 15% more information than the percentage of secondary structure.

From all proteins studied here, globin-like proteins are the most folded ones, e.g., the highest folding degree index is that of the hemoglobin i (homodimer) from Ark clam with PDB code 3sdh, which is a globin-like protein. Among the 10 most folded proteins studied here, 6 belong to the family

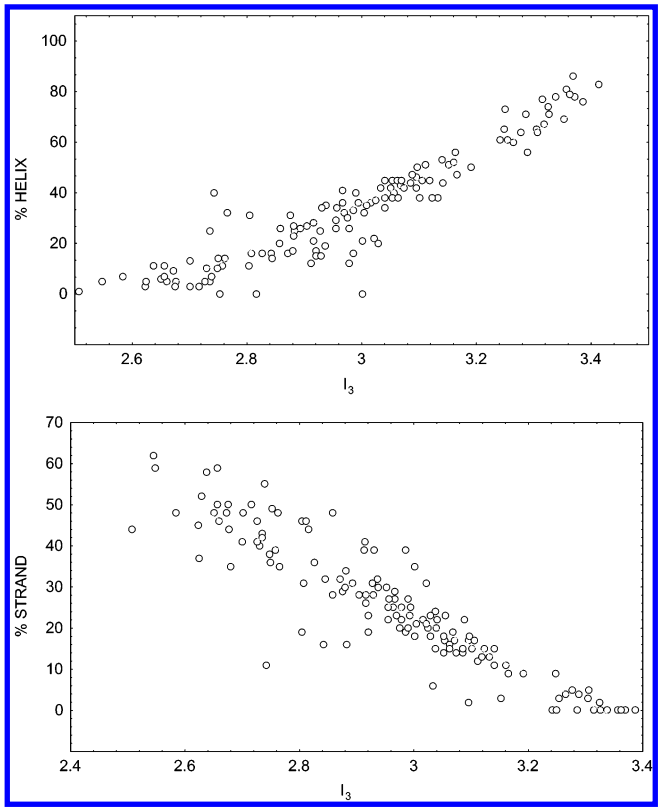

Figure 6. Dependence of the folding degree on composition of secondary structure measured as percentage of helix and strand.

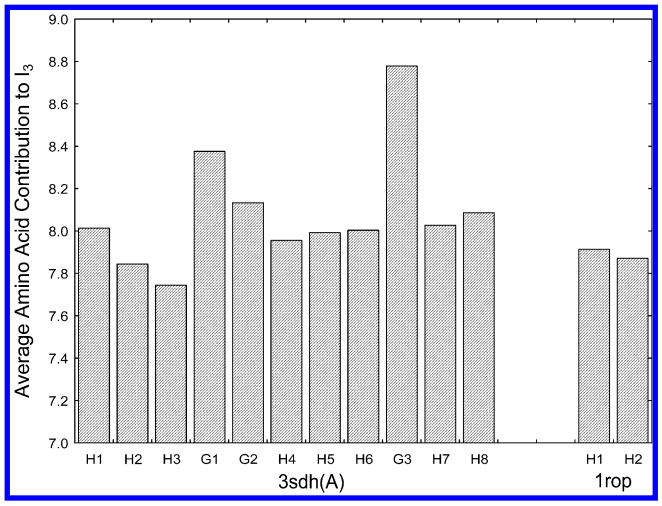

Figure 7. The folding degree of the different helices (H: α-helix; G: 3_{10} helix) of the most folded protein among the 152 nonhomologous proteins studied (3sdh(A)). Note that the values of amino acid contributions to the folding degree used for the averages have not been normalized by the number of residues in the protein to allow interprotein comparisons.

of globin-like proteins, which have a 6 helices, folded leaf, partially opened core. Those studied here contain two or three 3_{10} helices in their structures (the only one exception is C-phycocyanin, 1cpc, a globin-like protein that belongs to the class of light harvesting proteins that does not contain 3_{10} helices). We have calculated the amino acid contribution to the folding degree of 3sdh(A) protein and obtained the average values of the folding degree contribution of each helix in this protein. This is shown in Figure 7 where we have included the two alpha helices of the protein 1rop (a typical mainly alpha protein) for comparison. As can be seen, most of the helices in the globin-like protein have high folding degrees with the greatest contributions coming from the 3_{10} helices. Globin-fold comprises diverse sequences such as myoglobin, hemoglobin, and phycocyanin, and this is a good example of divergent evolution with a well-preserved three-dimensional structure of the proteins.

It is generally observed that the folding degree is well conserved in protein families. In Figure 8 we illustrate the conservation of our folding degree index for five different families of proteins. This observation is in agreement with that made by Chothia and Lesk[11] namely that the three-dimensional structure of proteins is more conserved in evolution than sequence. In consequence, the folding degree of protein families does not have to show a great variability in order to guarantee the three-dimensional structure necessary for protein function, e.g., optimal separation in space of residues in active sites.

I_3 AND DOMAIN STRUCTURAL CLASS ASSIGNMENT

Proteins can share a common fold in families and superfamilies if they have the same major secondary structures in the same arrangement with the same topological connections.[61,62] Each fold group is then grouped into one of the general classes. There are several databases that use this type of classification of proteins, such as SCOP (http://scop.mrc-lmb.cam.ac.uk/scop/)[63,64] and CATH (http://www.biochem.ucl.ac.uk/bsm/cath/).[65,66] The major classes in SCOP are mainly α, mainly β, α/β, and α + β. The difference between α/β and α + β is only based on the separation of helices and strands in the protein. In the first attempt to differentiate between these four classes, our folding degree index was unable to distinguish between α/β and α + β classes. The same result was obtained by Michie et al.,[67] who concluded that "the previously separate α/β and α + β classes show considerable overlap and are more naturally represented as a single α/β class". This conclusion was obtained after the analysis of 197 manually classified, nonhomologous domains. In fact, in the CATH database these two groups are joined together (α-β) forming only three major classes.[65,66] Both databases consider small proteins or domains of small secondary structure content, which often have little secondary structure and are held together by disulfide bridges or ligands, as a separate class. As the classification of proteins in such classes depends on their secondary structure, the folding degree index defined here is anticipated to discriminate well amongst these classes.

We have considered 100 single domain proteins from our data set that were classified by using CATH into one of the general classes. Small proteins were excluded from this classification as only 9 proteins were included in the data set, and they mainly consisted of a little secondary structure. Linear discriminant analysis was used to generate a classification model for the remaining 91 proteins. A general good classification of 89% was obtained for these proteins, 87% of mainly α (20/23), 90% of mainly β (37/41), and 88% of α-β proteins (21/24) were well classified. There were some proteins, such as 3il8, 9rnt, and 1cew, which are classified by CATH as mainly β but that according to their values of our folding degree index should correspond to the α-β class. These three proteins are classified by SCOP as α + β proteins, and we have considered them in the α-β

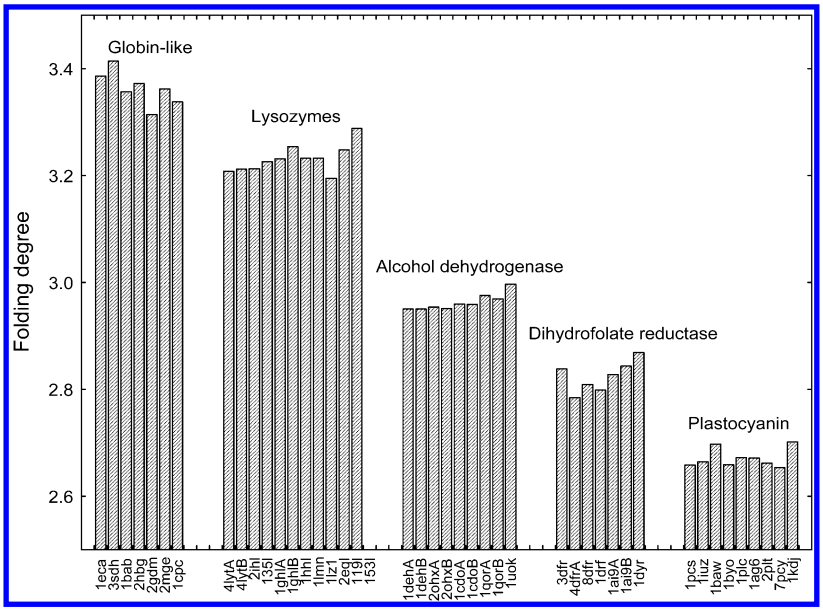

Figure 8. The folding degree of protein families.

class for the training classification. A similar situation occurs for 2cba, which is classified by CATH as α-β but as mainly β in SCOP (in agreement with our classification based on the folding degree index). The classification of three proteins was uncertain as the percentages of classification in two different classes do not differ by more than 10% (2end, 1dsb, and 2aza). The cutoffs values of I_3 for the proteins classes are as follows: mainly β, $I_3 < 2.8186$; α-β, $2.8330 < I_3 < 3.1512$; mainly α, $I_3 > 3.1658$. The values in the borderline correspond to uncertain regions where classification should be carried out by other means.

A test set selected from the work of Michie et al.[67] has been used to establish the utility of our folding degree index for an automated class assignment of proteins. After excluding coils (1bbt4, 1aaf, 1bal, 1pdc, 1rip), proteins with uncertainly assigned classes by the visual protocol (6inse, 1tabi, 1pyaa, 1oma), small proteins with little secondary structure (1d66(A)), proteins determined by NMR (1bha, 1lab, 1bw3, 1vil), (these proteins were removed to avoid a new source of error as we are only concerned here with proteins studied by X-ray diffraction), proteins wrongly assigned to other classes according to both CATH and SCOP (1pkp that contains two domains instead of being α-β and 1stf(I), which is mainly β instead of α-β), and proteins included in our training set (1lis, 1lts, 1huw, 1tml, 2cpl, 1cew(I)) a test set of 21 proteins remains. Nineteen of them are correctly classified by our folding degree index into any of the three classes, the classification of one is uncertain, and only one is misclassified resulting in a general good classification of 95% (19/20). As can be seen in Table 5 these results show the utility of the folding degree index introduced here for the automated class assignment of proteins. This index can be used together with other criteria, such as those currently in use, for classifying proteins into classes in a fully computerized way.

Table 5. Automatic Classification of Domains into Classes According to Their Folding Degree for a Test Set of Proteins

mainly α

PDB code	I_3	% class (α)	class
1hyp	3.2219	86.3	α
1rfb(A)	3.3443	99.4	α
1ysa(C)	3.3331	99.2	α
1acp	3.1445	43.0	α-β
1bgc	3.3346	99.3	α
1rib	3.2876	97.4	α

α-β

PDB code	I_3	% class (α-β)	class
5fd1	3.1317	65.3	α-β
2dnj(A)	2.8993	86.1	α-β
1mat	2.8962	85.1	α-β
1pya(B)	2.8571	67.4	α-β
1hge(B)	2.9836	97.4	α-β
3mon	2.8248	47.1	U

β

PDB code	I_3	% class (β)	class
2ltnb	2.4347	99.9	β
4htci	2.6079	99.7	β
1lya(A)	2.6186	99.6	β
2bpa2	2.5248	99.9	β
3mon(A)	2.4929	99.9	β
1tfi	2.6213	99.6	β
1cau(B)	2.7286	93.3	β
1shg	2.6056	99.7	β
1pnj	2.7729	81.3	β

CONCLUSIONS

We have shown how the intuitive idea of the folding degree can be formalized in a way that allows quantitative characterization. This quantitative measure enables ordering among protein chains based on their degree of folding, i.e., it allows the use of expression "more folded than" in comparing structures of proteins in a quantitative way. The current characterization of the folding degree also permits different types of helices to be distinguished, such as α, 3_{10}, and π as well as among the same type of helix in different environments.

The study of a representative data set of nonhomologous proteins shows a variability in the folding degree of about 27%. We have shown that external factors, such as crystal-

lization artifacts or experimental conditions (resolution, refinement protocol, temperature, pH), are not responsible for this variability as they account for less than 4% of it. Thus, our folding degree index mainly characterizes the intrinsic nature of folding in protein chains. As a result of this finding, our folding degree index shows a good linear relationship with the percentages of secondary structure in the protein and is able to account for the small changes produced in the structure due to crystal packing and temperature. This index has also shown good potential in applications such as the automatic classification of proteins into classes, such as mainly-α, mainly-β, and α- β.

ACKNOWLEDGMENT

The author thanks Ministerio de Ciencia y Tecnología, Spain for partial financial support under Ramón y Cajal Program.

Supporting Information Available: Protein databank (PDB) codes (fifth digit corresponds to the chain) and structural indices for the 152 nonhomologous proteins studied (Table 1). This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Roix, J.; Misteli, T. Genomes, proteomes, and dynamic networks in the cell nucleus. *Histochem. Cell Biol.* **2002**, *118*, 105−116.
(2) Gerling, I. C.; Solomon, S. S.; Bryer-Ash, M. Genomes, transcriptomes, and proteomes. Molecular medicine and its impact on medical practice. *Arch. Internal Med*. **2003**, *163*, 190−198.
(3) Rost, B. Did evolution leap to create the protein universe? *Curr. Opin. Struct. Biol*. **2002**, *12*, 409−416.
(4) Martí-Renom, M. A.; Stuart, A. C.; Fiser, A.; Sánchez, R.; Melo, F.; ×b3ali, A. Comparative protein structure modeling of genes and genomes. *Annu. Rev. Biophys. Biomol. Struct.* **2000**, *29*, 291−325.
(5) Heinemann, U.; Illing, G.; Oschkinat, H. High-throughput three-dimensional protein structure determination. *Curr. Opin. Biotech*. **2001**, *12*, 348−354.
(6) Montelione, G. T. Structural genomics: an approach to the protein folding problem. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98*, 13488−13489.
(7) Pilizota, T.; Lučić, B.; Trinajstić, N. Use of variable selection in modeling the secondary structural content of proteins from their composition of amino acid residues. *J. Chem. Inf. Comput. Sci*. **2004**, *44*, 113−121.
(8) Dill, K. A. Dominant forces in protein folding. *Biochemistry* **1990**, *29*, 7133−7155.
(9) Di Francesco, P. Folding and coloring problems in mathematics and physics. *Bull. Am. Math. Soc.* **2000**, *37*, 251−307.
(10) Soss, M.; Toussaint, G. T. Geometric and computational aspects of polymer reconfiguration. *J. Math. Chem*. **2000**, *27*, 303−318.
(11) Chothia, C.; Lesk, A. M. The relation between the divergence of sequence and structure in proteins. *EMBO J.* **1986**, *5*, 823−826.
(12) Vendruscolo, M.; Zurdo, J.; MacPhee, C. E.; Dobson, C. M. Protein folding and misfolding: A paradigm of self-assembly and regulation in complex biological systems. *Philos. Trans. Royal Soc. London, Ser A: Math., Phys. Eng. Sci*. **2003**, *361*, 1205−1222.
(13) Baldwin, M. A.; James, T. L.; Cohen, F. E.; Prusiner, S. B. The three-dimensional structure of prion protein: implications for prion disease. *Biochem. Soc. Trans*. **1998**, *26*, 481−486.
(14) Chu, R.; Pei, W.; Takei, J.; Bai, Y. Relationship between the Native-State Hydrogen Exchange and Folding Pathways of a Four-Helix Bundle Protein. *Biochemistry* **2002**, *41*, 7998−8003.
(15) Vishnyakov, A.; Neimark, A. V. Molecular simulation study of Nafion membrane solvation in water and methanol. *J. Phys. Chem*. B **2000**, *194*, 4471−4478.
(16) Nowick, J. S.; Cary, J. M.; Tsai, J. H. A Triply Templated Artificial β-Sheet. *J. Am. Chem. Soc*. **2001**, *123*, 5176−5180.
(17) dos Santos, D. J. V. A.; Gomes, J. A. N. F. Molecular dynamics study of a hexadecyltrimethylammonium chloride monolayer at the interface between two immiscible liquids. *Langmuir* **2003**, *19*, 958−966.
(18) Nowick, J. S.; Smith, E. M.; Ziller, J. W.; Shaka, A. J. Three-stranded mixed artificial β-sheets. *Tetrahedron* **2002**, *58*, 727−739.
(19) Hofman, M. A. Size and shape of the cerebral cortex in mammals. I. The cortical surface. *Brain Behav. Evol*. **1985**, *27*, 28−40.
(20) Gómez, B. J. Structure and functioning of the reproductive system. In *The Biology of Terrestrial Molluscs*; Baker, G. M., Ed.; CAB Int.: Wallingford 2001; pp 307−330.
(21) Batchelor, Ph. G.; Castellano Smith, A. D.; Hill, D. L. G.; Hawkes, D. J.; Cox, T. C. S.; Dean, A. F. Measures of folding applied to the development of the human fetal brain. *IEEE Trans. Medical Imag*. **2002**, *21*, 953−65.
(22) Estrada, E. Characterization of 3D molecular structure. *Chem. Phys. Lett*. **2000**, *319*, 713−718.
(23) Estrada, E. Characterization of the folding degree of proteins. *Bioinformatics* **2002**, *18*, 697−704.
(24) Estrada, E. Application of a novel graph-theoretic folding degree index to the study of steroid-DB3 antibody binding affinity. *Comput. Biol. Chem*. **2003**, *27*, 305−313.
(25) Estrada, E. Characterisation of the amino acids contribution to the folding degree of proteins. *Prot.: Struct. Funct. Bioinform*. **2004**, *54*, 727−737.
(26) *The English Oxford Dictionary*, 2nd ed.; Simpson, J. A., Weiner, E. S., Eds.; Oxford University Press: 1989.
(27) Gross, J. T.; Yellen, J. *Graph Theory and Its Applications*; CRC Press: Boca Raton, Fl, 1999; p 13.
(28) Aloupis, G.; Demaine, E. D.; Dujmović, V.; Erickson, J.; Langerman, S.; Meijer, H.; O'Rourke, J.; Overmars, M.; Soss, M.; Streinu, I.; Toussaint, G. T. Flat-state connectivity of linkages under dihedral motions. In *Lecture Notes in Computer Science;* Vancouver, British Columbia, Canada, Vol. 2518, pp 369−380.
(29) Weisstein, E. W. Dihedral angles. In *MathWorld. A Wolfram Web Resource*. http://mathworld.wolfram.com/dihedralangle.html.
(30) Demaine, E. D.; Langerman, S.; O'Rourke, J. Geometric restrictions on producible polygonal protein chains. In *Proc. 14th Int. Symp. Algor. Comput.* (*ISAAC 2003*). Kyoto, Japan, Dec. 15−1, 2003, in press. (http://theory.lcs.mit.edu/~edemaine/papers/.
(31) Estrada, E. Generalized Spectral Moments of the Iterated Line Graphs Sequence. A Novel Approach to QSPR Studies. *J. Chem. Inf. Comput. Sci*. **1999**, *39*, 90−95.
(32) Biggs, N. *Algebraic Graph Theory*; Cambridge University Press: Cambridge, 1993.
(33) Estrada, E. spectral moments of the edge-adjacency matrix of molecular graphs. 2. Molecules containing heteroatoms and QSAR applications. *J. Chem. Inf. Comput. Sci*. **1997**, *37*, 320−3278.
(34) Estrada, E.; Gutierrez, Y. Modeling chromatographic parameters by a novel graph theoretical sub-structural approach. *J. Chromatogr. A* **1999**, *858*, 187−199.
(35) Randić, M.; Krilov, G. Characterization of 3-D sequences of proteins. *Chem. Phys. Lett*. **1997**, *272*, 115−119.
(36) Randić, M.; Krilov, G. On characterization of the folding of proteins. *Int. J. Quantum Chem*. **1999**, *75*, 1017−1026;
(37) Bytautas, L.; Klein, D. J.; Randić, M.; Pisanski, T. Foldedness in linear polymers: A difference between graphical and Euclidean distances. *DIMACS Series Discuss. Math*. **2000**, *51*, 39−61.
(38) Eigenvalues and eigenvectors of a tridiagonal matrix. In *Numerical Recipes in FORTRAN 77: The Art of Scientific Computing*; Cambridge University Press. 1986−1992; pp 469−475.
(39) Gu, M.; Eisenstat, C. A divide-and-conquer algorithm for the symmetric tridiagonal eigenproblem. *SIAM J. Mater. Anal. Appl*. **1995**, *16*, 172−191.
(40) Fleming, P. J.; Richards, F. M. Protein packing: dependence on protein size, secondary structure and amino acid composition. *J. Mol. Biol*. **2000**, *299*, 487−498.
(41) Weisstein, E. W. Packing. In: *MathWorld. A Wolfram Web Resource*. http://mathworld.wolfram.com/packing.html.
(42) Liang, J.; Dill, K. A. Are proteins well-packed? *Biophys. J*. **2001**, *81*, 751−766.
(43) Pattabiraman, N.; Ward, K. B.; Fleming, P. J. Occluded molecular surface: analysis of protein packing. *J. Mol. Recog*. **1995**, *8*, 334−44.
(44) Flory, P. J. *Principles of Polymer Chemistry*; Cornell University Press: Ithaca, NY, 1953.
(45) Jacobson, M. P.; Friesner, R. A.; Xiang, Z.; Honig, B. On the role of the crystal environment in determining protein side-chain conformations. *J. Mol. Biol*. **2002**, *320*, 597−608.
(46) Carugo, O.; Argos, P. Protein−protein crystal-packing contacts. *Protein Sci*. **1997**, *6*, 2261−2263.
(47) Dasgupta, S.; Iyer, G. H.; Bryant, S. H.; Lawrence, C. E.; Bell, J. A. Extent and nature of contacts between protein molecules in crystal lattices and between subunits of protein oligomers. *Proteins* **1997**, *28*, 494−514.

(48) Ponsting, H.; Henrick, K.; Thornton, J. M. Discriminating between homodimeric and monomeric proteins in the crystalline state. *Prot.: Struct., Funct., Genet*. **2000**, *41*, 47−57.
(49) Valdar, W. S. J.; Thornton, J. M. Conservation helps to identify biologically relevant crystal contacts. *J. Mol. Biol*. **2001**, *313*, 399−416.
(50) Jones, S.; Thornton, J. M. Principles of protein−protein interactions. *Proc. Natl. Acad. Sci. U.S.A.* **1996**, *93*, 13−20.
(51) Driessen, H.; Haneef, M. I. J.; Harris, G. W.; Howlin, B.; Khan, G.; Moss, D. S. RESTRAIN: restrained structure-factor least-squares refinement program for macromolecular structures. *J. Appl. Crystallogr*. **1989**, *22*, 510−16.
(52) Tronrud, D. E.; Ten, E.; Lynn, F.; Matthews, B. W. An efficient general-purpose least-squares refinement program for macromolecular structures. *Acta Crystallogr*. **1987**, *A43*, 489−501.
(53) Brünger, A. T. X-PLOR, Versión 3.1. Yale University, New Haven, CT, 1992.
(54) Finzel, B. C. Incorporation of fast Fourier transforms to speed restrained least-squares refinement of protein structures. *J. Appl. Crystallogr*. **1987**, *20*, 53−5.
(55) Berisio, R.; Sica, F.; Lamzin, V. S.; Wilson, K. S.; Zagari, A.; Mazzarella, L. Atomic resolution structures of ribonuclease A at six pH values. *Acta Crystallogr*. **2002**, *D58*, 441−450.
(56) Tilton, R. F., Jr.; Dewan, J. C.; Petsko, G. A. Effects of temperature on protein structure and dynamics: X-ray crystallographic studies of the protein ribonuclease-A at nine different temperatures from 98 to 320K. *Biochemistry* **1992**, *31*, 2469−2481.
(57) Privalov, P. L. Thermodynamic problems of protein structure. *Annu. Rev. Biophys. Biophys. Chem.* **1989**, *18*, 47−69.
(58) Dengler, U.; Siddiqui, A. S.; Barton, G. J. Protein structural domains: analysis of the 3Dee domains database. *Prot.: Struct., Funct., Genet*. **2001**, *42*, 332−344.
(59) George, R. A.; Heringa, J. An analysis of protein domain linkers: Their classification and role in protein folding. *Protein Eng*. **2002**, *15*, 871−879.
(60) Fersh, A. *Structure and Mechanism in Protein Science*; N. H. Freeman & Co.: New Cork, 1999; p 18.
(61) Levitt, M.; Chothia, C. Structural patterns in globular proteins. *Nature* **1976**, *261*, 552−558.
(62) Wolf, Y. I.; Grishin, N. V.; Koonin, E. V. Estimating the Number of Protein Folds and Families from Complete Genome Data. *J. Mol. Biol*. **2000**, *299*, 897−905.
(63) Murzin, A. G.; Brenner, S. E.; Hubbard, T.; Chothia, C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol*. **1995**, *247*, 536−40.
(64) Lo Conte, L.; Brenner, S. E.; Hubbard, T. J. P.; Chothia, C.; Murzin, A. G. SCOP database in 2002: refinements accommodate structural genomics. *Nucl. Acids Res*. **2002**, *30*, 264−267.
(65) Orengo, C. A.; Michie, A. D.; Jones, S.; Jones, D. T.; Swindells, M. B.; Thornton, J. M. CATH − a hierarchic classification of protein domain structures. *Structure* **1997**, *5*, 1093−1108.
(66) Pearl, F. M. G.; Lee, D.; Bray, J. E.; Sillitoe, I.; Todd, A. E.; Harrison, A. P.; Thornton, J. M.; Orengo, C. A. Assigning genomic sequences to CATH. *Nucl. Acids Res*. **2000**, *28*, 277−282.
(67) Michie, A. D.; Orengo, C. A.; Thornton, J. M. Analysis of domain structural class using an automated class assignment protocol. *J. Mol. Biol*. **1996**, *262*, 168−185.

CI034278X