

Flexibility of α -Helices: Results of a Statistical Analysis of Database Protein Structures

Eldon G. Emberly, Ranjan Mukhopadhyay, Ned S. Wingreen and Chao Tang*

NEC Research Institute
4 Independence Way
Princeton, NJ 08540, USA

α -Helices stand out as common and relatively invariant secondary structural elements of proteins. However, α -helices are not rigid bodies and their deformations can be significant in protein function (e.g. coiled coils). To quantify the flexibility of α -helices we have performed a structural principal-component analysis of helices of different lengths from a representative set of protein folds in the Protein Data Bank. We find three dominant modes of flexibility: two degenerate bend modes and one twist mode. The data are consistent with independent Gaussian distributions for each mode. The mode eigenvalues, which measure flexibility, follow simple scaling forms as a function of helix length. The dominant bend and twist modes and their harmonics are reproduced by a simple spring model, which incorporates hydrogen-bonding and excluded volume. As an application, we examine the amount of bend and twist in helices making up all coiled-coil proteins in SCOP. Incorporation of α -helix flexibility into structure refinement and design is discussed.

© 2003 Elsevier Science Ltd. All rights reserved

*Corresponding author

Keywords: α -helices; database protein structures; protein folds

Introduction

Protein folds typically consist of two fundamental building blocks: α -helices and β -strands. These secondary elements pack together to form the final tertiary fold.^{1,2} However, the constraints of packing may be inconsistent with idealized conformations of the helices and strands. To what extent are these elements flexible?

One measure of flexibility is provided by a Ramachandran plot of the probability distribution of backbone dihedral angles.^{1,3} In such a plot, α -helices appear as a high-probability peak around $\phi = -50$, $\psi = -50$, while β -strands form a more diffuse peak around $\phi = -120$, $\psi = 120$. However, the flexibility of helices and strands is due to collective motion of many residues, and cannot be adequately described by the distribution of single $\{\phi, \psi\}$ pairs.

Collective deformations have been considered before in many biological contexts. Normal mode

analysis of protein structure has been performed to extract the flexible modes of proteins.^{4–10} The flexible motions identified in this way sometimes correspond to functional motions that the protein can perform.¹⁰ The flexibility of double-stranded DNA plays an important role both in the packing of DNA,¹¹ and in regulation of gene transcription *via* protein–DNA interactions.^{12–14} Recently, a principal-component analysis of database DNA structures was used to characterize the average deformation and deformability of all dinucleotide pairs.¹⁵

We have employed a similar principal-component analysis (PCA) to quantify the flexibility of α -helices. Helices with lengths ranging from $L = 10$ –25 residues were extracted from a representative set of α , $\alpha + \beta$, and α/β folds from the Protein Data Bank. We found that there are three dominant modes of flexibility: two nearly degenerate bend modes and one twist mode. It is natural to identify these as the three lowest normal modes of an α -helix, in particular since the distribution of amplitudes is consistent with independent Gaussians. According to elasticity theory, these low-lying normal modes should be insensitive to details of the interaction potential. Indeed, we found that a spring model with only two parameters reproduced not only the main bend and twist modes, but many higher-order modes as well.

Present address: E. G. Emberly, Center for Studies in Physics and Biology, The Rockefeller University, 1280 York Ave, New York, NY 10021, USA.

Abbreviation used: PCA, principal-component analysis.

E-mail address of the corresponding author: tang@research.nj.nec.com

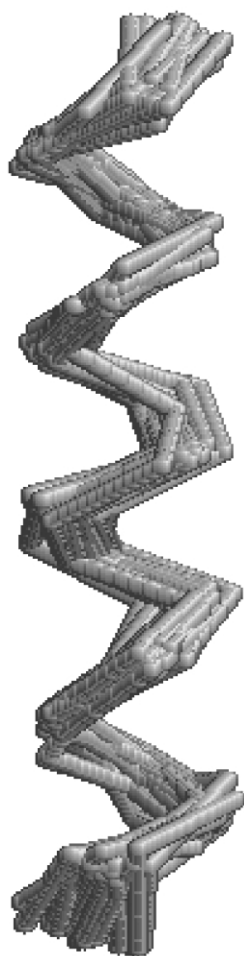


Figure 1. Representative set of 47 aligned 18-mer helices.

What relevance does α -helix flexibility have to biology? An obvious case is the formation of coiled coils of α -helices. A coiled coil is a domain of two or more α -helices wound around each other to form a superhelix. The α -helices typically interact with each other *via* buried hydrophobic residues, salt bridges, and interlocked side-chains.^{16,17} Such superhelical domains often contribute to protein-protein recognition, with helices from different proteins coming together to form the coiled coil. We have examined helices making up the known coiled coils. For relatively short coiled-coil helices we find that the deformations can be accounted for primarily by the bend and twist modes. For much longer helices, higher-order harmonics of bend are required to describe the deformation. In most of the cases examined, helices making up coiled coils can be well described using a minimal number of the lowest normal-modes of our spring model.

The quantification of α -helix flexibility may prove useful in structure refinement and protein design. For instance, folding studies that rely on rigid helical fragments may benefit from the inclusion of collective flexible motions to further optimize the energy of the sequence for the given

fold.^{18,19} A recent approach to protein design considers all possible packings of secondary structural elements;²⁰ however, so far only idealized rigid helices have been considered. Based on the current results, the low-lying bend and twist modes of helices can be incorporated to allow a more realistic balance between packing and deformation energies.

Results

Principal-component analysis of database helix structures

Sets of α -helices of given length were extracted from a representative set of protein structures (see Methods). To quantify the flexibility of these helices, we performed a structural principal-component analysis (PCA). For each length L , i.e. number of residues, we began by computing the mean helix structure *via* an iterative procedure. Starting with an ideal helix (periodicity 3.6 residues/turn, rise 1.5 Å/residue), we aligned the C^α positions of each length- L fragment in the representative set to the ideal helix[†]. A mean structure was then obtained by averaging the position of each C^α over these aligned structures. This procedure was then iterated, each time using the new, mean structure as the basis for the alignments, until the mean structure converged to within 10^{-4} Å/residue. An example of a set of aligned structures from the representative data set is shown in Figure 1 for helices of length $L = 18$.

The second step in the principal-component analysis was to compute the structural covariance matrix for each length L . The covariance matrix is a measure of correlations between coordinates. In our case, it is a square matrix of dimension $3L$ (three spatial dimensions for each of L C^α atoms), with element i, j defined as:

$$C_{i,j} = \frac{1}{N-1} \sum_{m=1}^N (x_{mi} - \langle x_i \rangle)(x_{mj} - \langle x_j \rangle) \quad (1)$$

where N is the number of helices of length L in the data set, x_{mi} is the i th coordinate of the m th structure, and $\langle x_i \rangle$ is the i th coordinate of the mean structure.

We then computed the eigenvalues, $\{\lambda_q\}$, and eigenvectors, $\{\vec{v}_q\}$, of the covariance matrix. The largest eigenvalues and corresponding eigenvectors represent the directions in the $3L$ dimensional space for which the data has the largest variance. These directions are the “soft modes” of the helices, i.e. those collective deformations that appear with largest amplitude in the data set. Figure 2(a) shows the top ten eigenvalues for helices of length $L = 18$. Each eigenvalue is given

[†] Each structure was aligned so that the coordinate root mean square (crms) distance between it and the mean structure was minimized.

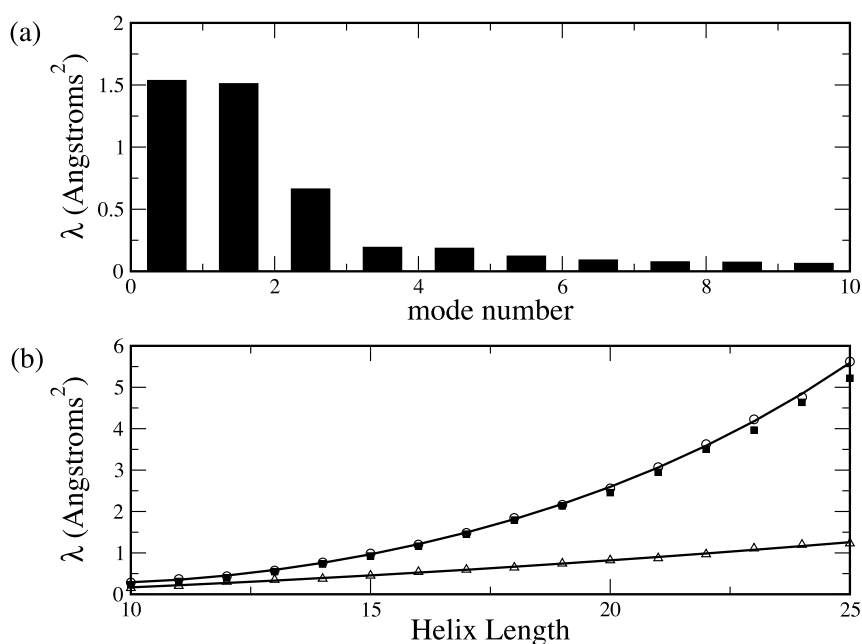


Figure 2. (a) The ten largest eigenvalues from the principal-component analysis of 18-residue α -helices from the representative data set. (b) The scaling of the three largest principal-component eigenvalues as a function of helix length L , i.e. number of residues. The bend modes are fit to the scaling form $\lambda_{\text{bend}} = (k_{\text{B}}T^*/\kappa)L^4$ yielding $k_{\text{B}}T^*/\kappa = 1.378 \times 10^{-5} \text{\AA}^{-2}$. The twist mode is fit to the scaling form $\lambda_{\text{twist}} = (k_{\text{B}}T^*/c)L^2$ yielding $k_{\text{B}}T^*/c = 0.0022$.

in units of \AA^2 and measures the variance of the distribution for a particular mode. Three dominant eigenvalues are evident in Figure 2(a). The first two modes are nearly degenerate and correspond to the bending of the helix in two orthogonal planes. This near degeneracy of the bend modes reflects the approximate cylindrical symmetry of an α -helix \dagger . The third mode is the overall twist of the helix. These dominant modes are shown with exaggerated amplitudes in Figure 3.

The scaling of the eigenvalues (i.e. variances) of the first three modes as functions of helix length is shown in Figure 2(b) for helices ranging from 10 to 25 residues. The eigenvalues of the bend modes grow with helix length approximately as L^4 while the eigenvalue for twist grows approximately as L^2 . This difference occurs because bend modes induce displacements from the mean helix structure which grow quadratically with helix length, while twist modes induce displacements which grow linearly with helix length. A model for the scaling of the eigenvalues based on an elastic rod is presented later in the paper.

Next, we look at the actual distribution of the data for the three dominant modes. For each helix fragment its displacement vector $\delta\vec{x} = \vec{x} - \langle\vec{x}\rangle$ can be expanded in terms of the eigenvectors giving $\delta\vec{x} = \sum_q a_q \vec{v}_q$. The amplitude a_q is given by the projection of the displacement vector onto mode q . Figure 4(a)–(c) shows histograms of the projections onto the two bend and one twist mode for helices

of length $L = 18$. For both the two bend modes and the one twist mode, the data have a nearly ideal Gaussian distribution. Best χ^2 -fits to Gaussians are shown by continuous lines. By definition, the exact variance of each distribution is equal to the eigenvalue λ_q for that mode. The variances of the best-fit Gaussians, 1.55 \AA^2 , 1.53 \AA^2 , and 0.71 \AA^2 , for the two bend and one twist mode, respectively, agree well with the exact variances, 1.53 \AA^2 , 1.51 \AA^2 , and 0.66 \AA^2 \ddagger .

By construction, the modes derived from the principal-component analysis are uncorrelated to

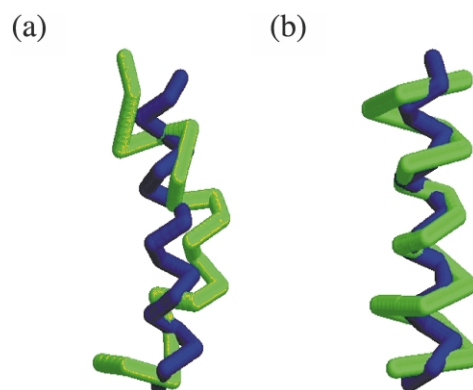


Figure 3. (a) Exaggerated bend mode of a helix (average structure in blue, bent structure in green). (b) Exaggerated twist mode of a helix (average structure in blue, twisted structure in green). The helices are 18 residues long.

\dagger If α -helices were actually cylinders, then our alignment procedure would produce only a single bend mode, since all the cylinders could be aligned to bend in a single direction. However, the positions of the C^α atoms of an α helix do not have cylindrical symmetry, so our alignment based on C^α positions gives the direction of bending of each helix unambiguously.

\ddagger The fitted values of the variances depend on the binning of the data. The shown binning yielded results that agree best with the exact variances of the distributions.

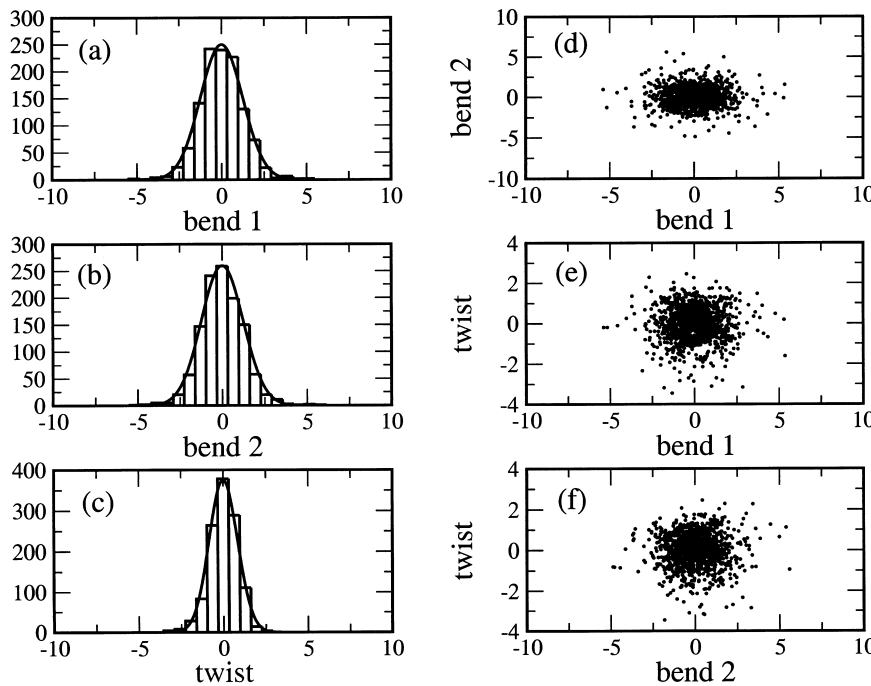


Figure 4. (a)–(c) Histograms of projections onto the two bend modes and one twist mode obtained from the principal-component analysis. Data are shown for the 1182 18-residue α -helices from the representative data set. Continuous lines correspond to Gaussian fits to the data. The fitted variances are 1.55 \AA^2 , 1.53 \AA^2 , and 0.71 \AA^2 , respectively, for the two bend modes and one twist mode. (d)–(f) Projections onto two-dimensional subspaces spanned by the two bend modes and one twist mode, for the same 1182 18-residue α -helices. The results are consistent with uncorrelated modes.

lowest order. That is, the expectations $\langle a_q a_{q'} \rangle$ are all zero, where a_q and $a_{q'}$ are amplitudes of different modes for a single helix. (This is simply the statement that the covariance matrix is diagonal in the basis of the eigenmodes.) However, there is no guarantee that the modes are uncorrelated at higher order. To look for possible correlations, we made scatter plots of the amplitudes of the three dominant modes in all pairwise combinations, shown in Figure 4(d)–(f) for the 1182 helices of length $L = 18$. The distributions of points in all three scatter plots are roughly ellipsoids with axes along x and y , indicating that there are no strong higher-order correlations between modes. This type of behavior was seen for all of the helix lengths $L = 10$ –25.

Normal-mode analysis of spring model for α -helices

While the dominant modes derived above come from studies of static structures, their properties are suggestive of *dynamical* normal modes. We show below that the two bend modes and one twist mode can be obtained from a simple model for the dynamics of a free α -helix. Moreover, the eigenvalue scaling and the uncorrelated Gaussian form of the distributions are characteristics of modes in thermodynamic equilibrium.

From an energetic point of view, an α -helix retains its helical shape due to two primary interactions. The first is the backbone hydrogen-bonding interaction between residues i and $i + 4$. The second is the excluded volume interaction between backbone atoms. We model these two terms by springs connecting nearby C^α atoms of an ideal helix. Again, we take an ideal helix to have period-

icity 3.6 residues/turn and rise $1.5 \text{ \AA}/\text{residue}$. The potential energy for the spring-model of the helix is given by:

$$V = \sum_i \sum_{m=1,2,3,4} \frac{1}{2} K_m (|\vec{r}_i - \vec{r}_{i+m}| - d_{i,i+m}^0)^2 \quad (2)$$

where \vec{r}_i is the position of the i th C^α atom, and d_{ij}^0 is the equilibrium distance between the i th and j th C^α atoms. In equation (2), there are springs connecting pairs of residues up to four apart along the chain, and so there are four spring constants K_m for $m = 1, 2, 3, 4$. However, we consider the limit $K_1 \rightarrow \infty$ which holds nearest-neighbor C^α atoms at a fixed distance of 3.8 \AA , and we set $K_{2,3} = K_2 = K_3$, leaving only two spring constants, $K_{2,3}$ and K_4 , as parameters.

The normal modes of the model α -helix are obtained by diagonalizing the $3L \times 3L$ spring matrix:

$$V_{i,j} = \frac{\partial^2 V}{\partial x_i \partial x_j}, \quad (3)$$

where x_i is the i th member of the $3L$ coordinates describing the helix. The eigenvalues \tilde{K}_q determined from the normal-mode analysis represent effective spring constants for each of the normal modes. The matrix has six zero eigenvalues, corresponding to the six rigid-body rotations and translations, for which there are no return forces. The lowest non-zero eigenvalues are the lowest-energy normal modes of the helix. Over a broad range of values for the spring constants, the first two non-zero modes are bend modes and the third is twist for helices up to length 33 (beyond this length, higher order harmonics of bend occur before twist), consistent with the dominant modes of

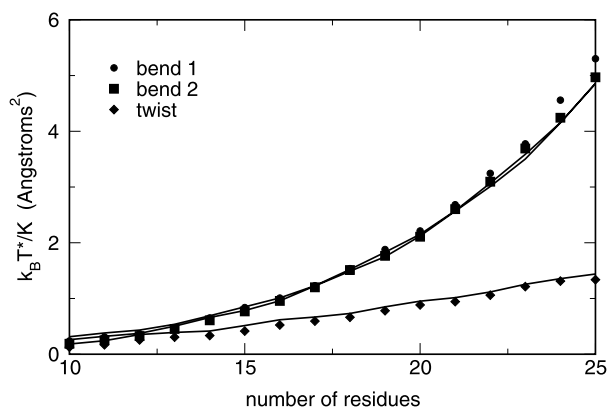


Figure 5. The three largest principal-component eigenvalues for helices of length $L = 10$ – 25 (discrete data points) compared to the inverse spring constants for the normal modes obtained from the spring model (continuous curves). To obtain this fit, we used real-space spring constants $K_{2,3} = 20k_B T^*/\text{\AA}^2$ and $K_4 = 7k_B T^*/\text{\AA}^2$.

static helices found from the PCA. Typically for the lengths studied, the top seven to ten modes from the normal-mode analysis agree very well with those obtained from PCA. At thermal equilibrium, these dynamical modes would follow the Boltzmann distribution $P(a_q) \approx \exp(-\tilde{K}_q a_q^2 / 2k_B T)$, where $P(a_q)$ is the probability of observing the q th mode with amplitude a_q and $1/2\tilde{K}_q a_q^2$ is the potential energy of the mode.

A more detailed comparison between the PCA and the spring model can be made by conjecturing that the Gaussian distributions of PCA modes represent *equilibrium* Boltzmann distributions at some effective temperature T^* . (Below, we discuss the use of T^* rather than room temperature T .) With this conjecture, one has the relation:

$$\exp\left(-\frac{a_q^2}{2\lambda_q}\right) = \exp\left(-\frac{\tilde{K}_q^{(\text{PCA})} a_q^2}{2k_B T^*}\right) \quad (4)$$

where the a_q are the mode amplitudes. So the effective spring constants of the PCA modes are given by $\tilde{K}_q^{(\text{PCA})} = k_B T^* / \lambda_q$. In other words, the PCA eigenvalues λ_q can be interpreted as inverse spring constants, with a proportionality constant $k_B T^*$, i.e. $\lambda_q = k_B T^* / \tilde{K}_q^{(\text{PCA})}$. Figure 5 shows a plot of the eigenvalues λ_q for the first three PCA modes, compared with $k_B T^* / \tilde{K}_q$ using the spring constants \tilde{K}_q for the first three low-energy modes determined from the normal-mode analysis of the spring model. The real-space spring constants that give this fit are $K_{2,3} = 20k_B T^* / \text{\AA}^2$ and $K_4 = 7k_B T^* / \text{\AA}^2$. The agreement between the PCA modes and the normal modes of the spring model is striking for both eigenvalues and eigenvectors (Figure 6). Note that there are only two free parameters in the spring model $K_{2,3}$ and K_4 , and that the mode shapes depend only on their ratio. Thus the dominant modes of static α -helices extracted from the

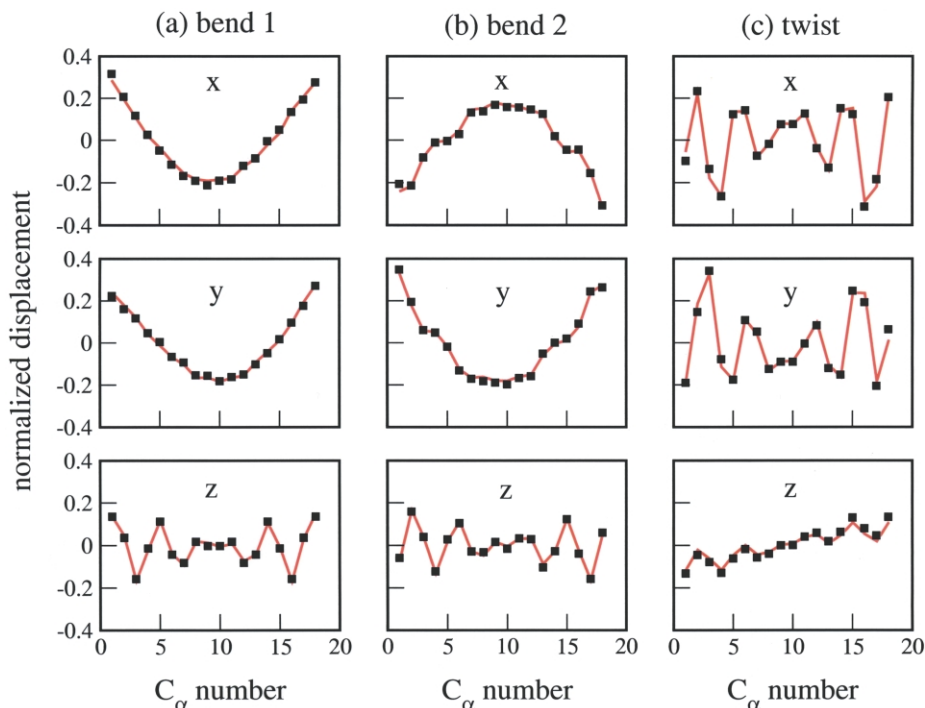


Figure 6. Plots of x , y , and z components of eigenvectors for the two bend modes, (a) and (b), and for the twist mode (c) for $L = 18$. Shown is the overlap of eigenvectors from the principal-component analysis (filled squares) with those from the spring model (red curves). The curves are not fits to the PCA eigenvector data; the curves are the eigenvectors from the spring model with the same spring constants used in Figure 5.

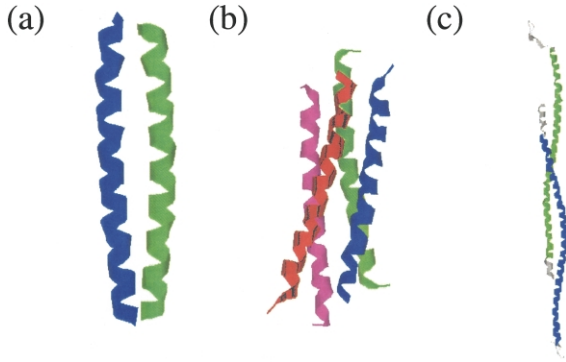


Figure 7. (a) Helices making up coiled coil in the leucine zipper (2ZTA). (b) Tetramerization domain of Mnt repressor (1QEY). (c) Coiled-coil fragment from chicken fibrinogen (1M1J).

database can be identified with the normal modes of simple spring model for a helix.

Scaling of the PCA modes

Guided by the interpretation of the PCA modes as normal modes, the scaling of the PCA eigenvalues can be understood relatively simply in terms of the bending and twisting of an elastic rod. For a uniformly bent rod, the displacement away from vertical goes as $\delta x \approx l^2/R$, where l is the length along the rod, and R is the radius of curvature. The normalized eigenvector describing this bending mode has the form $\vec{v} \sim (R^2/L^5)^{1/2} \times (L^2/R, \dots, L^2/R)$. Within a principal-component analysis, the eigenvalue for this bend mode is the average square of the projection of the displacement of the rod onto this mode. So the bend eigenvalue is given by:

$$\lambda_{\text{bend}} = \langle |\delta \vec{x} \cdot \vec{v}|^2 \rangle \sim \left\langle \left| L \frac{L^4}{R^2} \frac{R}{L^{5/2}} \right|^2 \right\rangle = L^5 \left\langle \frac{1}{R^2} \right\rangle \quad (5)$$

At thermal equilibrium each normal mode has $k_B T/2$ of potential energy. For the bend mode, this energy is put into the curvature of the rod:

$$\frac{1}{2} k_B T = \frac{1}{2} \kappa L \left\langle \frac{1}{R^2} \right\rangle \quad (6)$$

Substituting this equilibrium relation for $\langle 1/R^2 \rangle$ into equation (5) for the bend eigenvalue gives:

$$\lambda_{\text{bend}} \sim L^5 \left\langle \frac{1}{R^2} \right\rangle = \frac{k_B T}{\kappa} L^4 \quad (7)$$

Thus, from thermodynamic arguments, we find that the principal-component eigenvalue of the bend mode of an elastic rod scales as L^4 , as was found in Figure (2) for the bend eigenvalue of α -helices.

For twist, we assume that the rod twists uniformly by an angle $\delta\theta$ per unit length. The displacement associated with twist along the rod is given by $\delta x \sim l\delta\theta$, and hence the normalized vector

describing this mode is $\vec{v} \sim 1/(\delta\theta^2 L^3)^{1/2} \times (-L\delta\theta, \dots, L\delta\theta)$. Using the same formulation as above, we find that the eigenvalue for the twist mode goes as:

$$\lambda_{\text{twist}} \sim \left\langle \left| L \frac{(\delta\theta^2 L^2)}{(\delta\theta^2 L^3)^{1/2}} \right|^2 \right\rangle = \langle \delta\theta^2 \rangle L^3 \quad (8)$$

At thermodynamic equilibrium, the energy associated with the twist mode is:

$$\frac{1}{2} k_B T = \frac{c}{2} \langle \delta\theta^2 \rangle L \quad (9)$$

where c is a spring constant associated with twist. Substituting this equilibrium result for $\langle \delta\theta^2 \rangle$ into equation (8) for the twist eigenvalue gives:

$$\lambda_{\text{twist}} \sim \langle \delta\theta^2 \rangle L^3 = \frac{k_B T}{c} L^2 \quad (10)$$

So, we find that the principal-component eigenvalue of the twist mode of an elastic rod scales as L^2 , as was found in Figure (2) for the twist eigenvalue of α -helices.

Thus, the eigenvalues of bend and twist extracted from the PCA of α -helices are seen to scale with length in the same manner as those of a fluctuating elastic rod at thermal equilibrium. The difference between the scaling exponents, L^4 for bend and L^2 for twist, can be traced to the length dependence of displacements. For bend modes, displacements grow quadratically with length, $\delta x \approx l^2/R$, while for twist modes displacements grow linearly, $\delta x \sim l\delta\theta$.

Application to helices forming coiled coils

In this section we examine to what degree helices making up coiled coils can be described using the lowest-energy normal modes. We created a database of 222 coiled-coil structures from the coiled coil classification in SCOP.²¹ We then extracted 680 helices from the above structures and performed the analysis described below. The results of the analysis are available as Supplementary Material. Here we have chosen to focus on three representative coiled-coil structures. The first is a leucine zipper (2ZTA), which consists of two interacting helices. The second is the tetramerization domain of Mnt repressor which is representative of coiled coils that form as a result of protein-protein interactions (1QEY). Lastly, we consider two long helices that form a part of a coiled coil in the structural protein fibrinogen from chicken (1M1J). The coiled coils that we have chosen to focus on are shown in Figure 7.

For each coiled-coil helix we computed the normal modes of an ideal helix of identical length using our spring model. We then aligned the coiled-coil helix to the ideal helix and computed the displacement vector $\delta \vec{x}$, which by definition can be expanded in terms of the spring-model normal-mode eigenvectors $\delta \vec{x} = \sum_q a_q \vec{v}_q$. We then projected $\delta \vec{x}$ onto each eigenvector yielding

Table 1. Results of projecting coiled-coil helices onto normal-modes of spring model

| PDB ID | Residues | L | Bend | Bend ⁽²⁾ | Bend ⁽³⁾ | Twist | Total |
|--------|----------|-----|-------|---------------------|---------------------|-------|-------|
| 2ZTA | A 1–30 | 30 | 75.68 | 4.20 | 3.97 | 3.31 | 87.16 |
| | B 1–30 | 30 | 73.18 | 0.99 | 0.00 | 17.59 | 91.76 |
| 1QEY | A 54–81 | 28 | 76.01 | 1.56 | 2.72 | 9.61 | 89.90 |
| | C 54–81 | 28 | 76.03 | 1.55 | 2.72 | 9.61 | 89.91 |
| | B 54–80 | 27 | 86.25 | 1.07 | 0.00 | 6.75 | 94.07 |
| | D 54–80 | 27 | 86.23 | 1.07 | 0.00 | 6.75 | 94.05 |
| 1M1J | A 82–161 | 80 | 61.00 | 16.77 | 3.49 | 10.23 | 91.49 |
| | B 85–164 | 80 | 1.18 | 65.36 | 1.68 | 17.74 | 85.96 |

Columns denote the percentage of the helical displacement accounted for by the specified mode. Bend is the sum of the percentages for the two principal bend modes. Bend⁽²⁾ and Bend⁽³⁾ correspond, respectively, to percentages for the first and second harmonics of bend. Twist is the percentage for the principal twist mode. L is the length of the coiled-coil helix. See Supplementary Material for analysis of 680 helices in coiled-coil structures.

projection amplitudes a_q . The percentage of the displacement vector due to a single mode q is given by $a_q^2 / \sum_q a_q^2$. In Table 1, we show the percentages of the coiled-coil helix displacements captured by the sum of the principal bend normal modes, Bend, the first and second harmonics of bend, Bend⁽²⁾ and Bend⁽³⁾ and lastly Twist (the complete Table is available as Supplementary Material). For the shorter helices in the leucine zipper and Mnt, we find that the helical displacements are described predominantly by the bend modes, with some twist. Thus, coiled coils that are formed by shorter helices can be described well using just the bend and twist modes of the spring model. For the larger helices making up fibrinogen, where there is clear evidence of supercoiling, the second harmonic of bend is required. The second helix of the fibrinogen coil (green helix in Figure 6(c)) has 68% of its displacement captured by the two second harmonics of bend. Thus helical supercoiling is captured by higher-order harmonics of the

fundamental bend mode. For the helices of length 80, the second and third harmonics of bend are lower in energy than the twist mode, so twist is no longer the third lowest normal mode for longer helices.

To gain further insight into the deformations that helices undergo in coiled coils we have plotted the average fraction of the deformation for each mode (Bend, Bend⁽²⁾, Bend⁽³⁾, and Twist) as a function of helix length in Figure 8. From above, we know that the deformations of a helix at thermal equilibrium are related to the effective spring constants Kq of the normal modes via $\lambda_q = k_B T^* / Kq$. Thus, from the eigenvalues of the spring-model normal modes we can estimate the fraction of the total deformation that comes from each mode at thermal equilibrium:

$$\text{fraction}_q = \frac{\lambda_q}{\sum_{q'} \lambda_{q'}}. \quad (11)$$

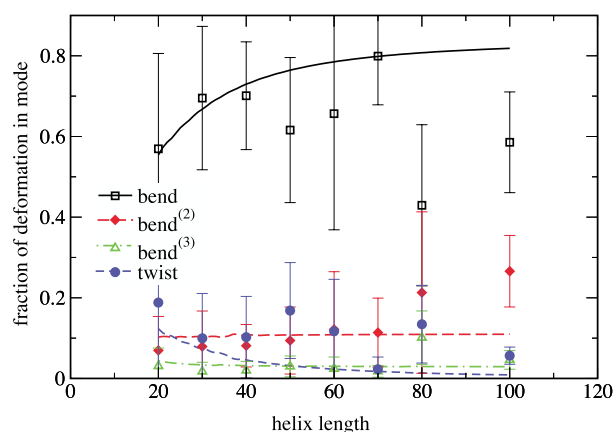


Figure 8. Average fraction of total deformation in top normal modes as a function of helix length for 680 helices making up coiled-coil structures: principal bend modes (open squares), first harmonics of bend modes (filled diamonds), second harmonics of bend modes (open triangles), and twist mode (filled circles). Curves correspond to expected average fractions for helices in thermal equilibrium.

In Figure 8, the actual fractions in the top modes for coiled coils are compared with the fractions expected at thermal equilibrium. For helix lengths < 60 the coiled-coil deformations are consistent with an equilibrium distribution. However, for longer helices, there is a departure from equilibrium in favor of the higher harmonics of the bend modes. This supports the conventional view that hydrophobic forces along with specific side-chain interactions cause super-coiling of helices in long coiled coils.

Discussion

Connection between static and dynamical modes of helices

Our principal-component analysis has shown that α -helices of lengths up to 25 residues, have three dominant independent “soft modes”: two bend and one twist. These modes were determined from static α -helix structures in the Protein Data Bank. The principal modes determined from these

static snap-shots agree extremely well with the dynamical normal modes of an α -helix obtained from a simple spring model. The projections of the static α -helices onto these three principal modes yield Gaussian distributions, which coincides with the distribution expected for dynamical equilibrium fluctuations. Why should an ensemble of static α -helical structures be related to the normal-mode fluctuations of a helix at thermal equilibrium? This connection can be understood if the ensemble of static α -helical structures has been sampled from a system that is under the influence of random forces. In a given protein structure, helices adopt conformations so that the forces acting on them add to zero. Over the entire ensemble of protein folds it is reasonable to assume that the forces that an α -helix experiences are approximately random. An elastic object acted on by random external forces is equivalent to that same object at thermal equilibrium at some effective temperature T^* . The fluctuations in the energy of the ensemble of α -helical structures set the effective temperature T^* . (Since the forces, or more precisely, energies involved in protein folding (hydrophobic interactions, hydrogen bonding, van der Waals etc.) all have the scale of the order of kcal/mol, or a few $k_B T$, with T being the room temperature, the effective temperature T^* obtained from PCA should be of the order of room temperature. Indeed, our fitted value of $K_4 = 7k_B T^*/\text{\AA}^2$, the hydrogen bond spring constant, is quite consistent with the hydrogen bond energy with T^* being room temperature.) The distribution of static helical structures sampled from a large ensemble of proteins will therefore have a distribution that is equivalent to a helix at thermal equilibrium at some temperature T^* . If the forces that helices experience within protein structures were systematically non-random then the resulting PCA distributions would depart from those of a helix at thermal equilibrium.

Incorporating helix bend and twist into models of protein structure

The results presented here can potentially be applied to structure refinement and structure design. Most off-lattice structure models of proteins fall into two classes: those with rigid secondary elements,^{19,22} or those in which every atom is free to move independently.^{23,24} The first has the advantage of locking out many of the degrees of freedom of the peptide chain. It has the disadvantage of potentially missing lower-energy conformations which could be accommodated if the secondary elements were flexible. The second approach, allowing every atom to move independently of the others, has the advantage that each atom is in principle allowed to find its equilibrium position within the fold. It has the great disadvantage of allowing all possible degrees of freedom, which greatly increases the complexity. A model which fits somewhere in between the

two extremes, allowing only a few important internal degrees of freedom, would be advantageous in many cases.

The dominant low lying normal modes of a helix can easily be incorporated into models of protein structure that currently use rigid helical segments. Each mode has an effective spring constant \tilde{K}_q , and eigenvector $\vec{v}_q = (x_{q,1}, x_{q,2}, \dots, x_{q,3L})$, which can be obtained by diagonalizing the spring matrix (3). The energy cost (in $k_B T^*$) for exciting these internal degrees of freedom is:

$$E = \sum_q \frac{1}{2} \tilde{K}_q a_q^2 \quad (12)$$

This prescription gives a simple way to include the internal degrees of freedom, along with the appropriate energy term, into models of protein structure. For shorter helices, only the two bend and twist modes need be incorporated. For longer helices that might supercoil, including higher order bend harmonics would be required. But nevertheless, describing the possible deformations of a helix can be described by adding relatively few extra degrees of freedom.

In summary, we have shown that α -helices have three prominent flexible modes: two bend and one twist. The principal modes obtained from static structures in the Protein Data Bank agree extremely well with the dynamical normal modes of a simple spring model of a helix. Moreover, the static α -helices from the database have independent Gaussian distributions of mode amplitudes, consistent with a quasi-thermal equilibrium. Use of these dominant soft modes may provide an intermediate path between rigid secondary-structures and independent all-atom models for protein structure refinement and design.

Methods

To compile a set of protein structures containing α -helices, we selected one representative of each fold in the α , $\alpha + \beta$, and α/β families from SCOP release 1.55,²¹ yielding a total of 399 protein structures. Each of these structures was then decomposed into its $\{\phi, \psi, \Omega\}$ angle sequence, and backbone bond lengths. All backbone atom coordinates could be reconstructed from this data. The $\{\phi, \psi, \omega\}$ angles for each of the 399 protein structures from SCOP were calculated using the freely available program Stride.²⁵

Sets of α -helices of given length were extracted from the above structure set as follows. We identified α -helices by unbroken series of dihedral angles within a square region $\{\phi, \psi = -50 \pm 30, -50 \pm 30\}$. For example, a sequence of 15 $\{\phi, \psi\}$ angles all falling within the defined region would be added to our helix set of 15mers. This same sequence would also contribute two 14mers, three 13mers, four 12mers, and so on, to the data sets of these other lengths. For a given helix length, we scanned all 399 structures, and extracted the α -helical fragments. This yielded our representative set of α -helices for lengths $L = 10-25$.

The eigenvalues and eigenvectors of both the covariance matrix and the spring matrix were computed using the eigenvalue solver for real symmetric matrices in the NAG numerical library. The elements making up the spring matrix, equation (3), were evaluated by computing the second derivative of equation (2) numerically.

Acknowledgements

We would like to thank David Moroz for rewarding discussions. One of us (N.S.W.) acknowledges valuable conversations with Bill Bialek and Arnold Neumaier.

References

- Richardson, J. S. (1981). The anatomy and taxonomy of protein structure. *Advan. Protein Chem.* **34**, 167–339.
- Chothia, C., Levitt, M. & Richardson, D. (1977). Structure of proteins: packing of α -helices and pleated sheets. *Proc. Natl Acad. Sci. USA*, **74**, 4130–4134.
- Ramachandran, G. N. & Sasisekharan, V. (1968). Conformations of polypeptides and proteins. *Advan. Protein Chem.* **28**, 283–437.
- Kidera, A. & Go, N. (1990). Refinement of protein dynamic structure: normal mode refinement. *Proc. Natl Acad. Sci. USA*, **87**, 3718–3722.
- Diamond, R. (1990). On the use of normal modes in thermal parameter refinement: theory and application to the bovine pancreatic trypsin inhibitor. *Acta Crystallog. sect. A*, **46**, 425–435.
- Faure, P., Micu, A., Perahia, D., Doucet, J., Smith, J. C. & Benoit, J. P. (1994). Correlated intramolecular motions and diffuse X-ray scattering in lysozyme. *Nature Struct. Biol.* **1**, 124–128.
- Tirion, M. M. (1996). Large amplitude elastic motions in proteins from a single-parameter, atomic analysis. *Phys. Rev. Letters*, **77**, 1905–1908.
- Haliloulu, T., Bahar, I. & Erman, B. (1997). Gaussian dynamics of folded proteins. *Phys. Rev. Letters*, **79**, 3090–3093.
- Krebs, W. G., Alexandrov, V., Wilson, C. A., Echols, N., Yu, H. & Gerstein, M. (2002). Normal mode analysis of macromolecular motions in a database framework: developing mode concentration as a useful classifying statistic. *Proteins: Struct. Funct. Genet.* **48**, 682–695.
- Bahar, I., Erman, B., Jernigan, R. L., Atilgan, A. R. & Covell, D. G. (1999). Collective motions in HIV-1 reverse transcriptase: examination of flexibility and enzyme function. *J. Mol. Biol.* **285**, 1023–1037.
- Travers, A. A. (1987). DNA bending and nucleosome positioning. *Trends Biochem. Sci.* **12**, 108–112.
- Anderson, J. E., Ptashne, M. & Harrison, S. C. (1987). Structure of the repressor-operator complex of bacteriophage 434. *Nature*, **326**, 846–852.
- Lewis, M., Chang, G., Horton, N. C., Kercher, M. A., Pace, H. C., Schumacher, M. A. *et al.* (1996). Crystal structure of the lactose operon repressor and its complexes with DNA and inducer. *Science*, **271**, 1247–1254.
- Schumacher, M. A., Choi, K. Y., Zalkin, H. & Brennan, R. G. (1994). Crystal structure of Lac I member, PurR, bound to DNA: minor groove binding by α -helices. *Science*, **266**, 763–770.
- Olson, W. K., Gorin, A. A., Lu, X.-J., Hock, L. M. & Zhurkin, V. B. (1998). DNA sequence-dependent deformability deduced from protein-DNA crystal complexes. *Proc. Natl Acad. Sci. USA*, **95**, 11163–11168.
- Crick, F. H. C. (1953). The packing of α -helices: simple coiled coils. *Acta Crystallog.* **6**, 689–697.
- Cohen, C. & Parry, D. A. D. (1986). Alpha-helical coiled coils—a widespread motif in proteins. *Trends Biochem. Sci.* **11**, 245–248.
- Simons, K. T., Bonneau, R., Ruczinski, I. & Baker, D. (1999). *Ab initio* protein structure prediction of CASP III targets using ROSETTA. *Proteins: Struct. Funct. Genet.* **37**, 171–176.
- Simons, K. T., Bonneau, R. & Baker, D. (2001). Prospects for *ab initio* protein structural genomics. *J. Mol. Biol.* **306**, 1191–1199.
- Emberly, E. G., Wingreen, N. S. & Tang, C. (2002). Designability of α -helical proteins. *Proc. Natl Acad. Sci. USA*, **99**, 11163–11168.
- Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia, C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **247**, 536–540.
- Park, B. H. & Levitt, M. (1995). The complexity and accuracy of discrete state models of protein structure. *J. Mol. Biol.* **249**, 493–507.
- Duan, Y. & Kollman, P. A. (1998). Pathways to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution. *Science*, **280**, 740–744.
- Lazaridis, T. & Karplus, M. (1997). “New View” of protein folding reconciled with the old through multiple unfolding simulations. *Science*, **278**, 1928–1931.
- Frishman, D. & Argos, P. (1995). Knowledge-base protein structure secondary structure assignment. *Proteins: Struct. Funct. Genet.* **23**, 566–579.

Edited by J. Thornton

(Received 26 September 2002; received in revised form 13 January 2003; accepted 15 January 2003)

SCIENCE @ DIRECT®
www.sciencedirect.com

Supplementary Material for this paper is available on Science Direct