A New Approach for Determining Low-Frequency Normal Modes in Macromolecules

PHILIPPE DURAND, GEORGES TRINQUIER, and YVES-HENRI SANEJOUAND

Laboratoire de Physique Quantique, IRSAMC-CNRS, Université Paul-Sabatier, 31062 Toulouse Cedex, France

SYNOPSIS

A new method for calculating a set of low-frequency normal modes in macromolecules is proposed and applied to the case of proteins. In a first step, the protein chain is partitioned into blocks of one or more residues and the low-frequency modes are evaluated at a low-resolution level by combining the local translations and rotations of each block. In a second step, these low-resolution modes are perturbed by high-frequency modes explicitly calculated in each block, thus leading to the exact low-frequency modes. The procedure is tested for three cases—decaalanine, icosaleucin, and crambin—using a perturbation-iteration scheme in the second step. Convergence properties and numerical accuracy are assessed and tested for various partitions. The low-resolution modes obtained in the first step are always found to be good starting approximations. Potential advantages of the method include a central processing unit time roughly N^2 dependent on the size of the problem (N being the number of degrees of freedom), the possibility of using parallel processing, the nonrequirement for loading the complete mass-weighted second-derivative input matrix into central memory, and the possibility of introducing in the procedure further structural hierarchy, such as secondary structures or motifs. In addition, any improvement or refinement of the algorithm benefits from the efficient formalism of the effective Hamiltonian theory. © 1994 John Wiley & Sons, Inc.

INTRODUCTION

In many proteins, the binding of some specific ligands induces large conformational changes. In some cases—hexokinase,[1,2] citrate synthase,[3,4] haemoglobin,[5,6] lysozyme,[7] etc.—both nonliganded and liganded structures have been determined through x-ray crystallography, allowing detailed analysis of the conformational changes occurring upon ligand binding. In the two-domain protein citrate synthase, substrate binding induces a 18° rotation of the small domain, closing the cleft between the two domains in which the substrate binding site lies, and providing a solvent-shielded environment for the catalysis to occur.[8,9]

In most cases, numerous atoms are involved in these large conformational changes. Such collective motions are difficult to study experimentally at the atomic level since information on motions of distant parts of the protein are simultaneously needed. As a matter of fact, little experimental data exists on the dynamics of "hinge-bending" motions. A number of theoretical analyses, however, have been made.[10,11] One of the best suited theoretical methods for studying collective motions in proteins is the normal mode analysis, which leads to the expression of the dynamics straightway from the superposition of collective variables, namely the normal mode coordinates.[12-14] Such a method has been successfully applied to the study of the lysozyme hinge-bending motion, where the lowest frequency normal mode (at around 3 cm^{-1}) was shown to have dominant contribution.[15-17] That low-frequency modes of motion (under 30–120 cm^{-1}, according to different criteria) are responsible for most of the amplitude of atomic displacements in proteins happens to be a general result.[15,18,19]

Several methods making use of low-frequency normal modes have been developed for studying pathways in conformational changes or finding saddle points.[20,21] In the Cerjan–Miller approach,[22] for

Biopolymers, Vol. 34, 759–771 (1994)
© 1994 John Wiley & Sons, Inc. CCC 0006-3525/94/060759-13

instance, starting from a local minimum, energy is maximized with respect to displacements along a given normal coordinate (chosen to be the most relevant throughout the process) while it is simultaneously minimized for the other normal coordinates, the process being repeated until a saddle point is reached. Applying such methods to the study of large conformational changes like the one recently analyzed in citrate synthase[23] would certainly lead to decisive insights in the understanding of mechanisms and kinetics of such motions.

At a more general level, normal coordinates have also proven to be powerful tools in the analysis of molecular dynamics trajectories through the quasiharmonic approximation,[24-26] or to sample the configurational space[27] when combined with the Monte Carlo method,[28] or as an aid to B-factor refinement in protein crystallography.[29] In other words, normal coordinates form a well-suited coordinate reference system for studying properties of the potential energy surfaces of macromolecules in general and of proteins in particular.

To date, the largest protein studied with normal modes analysis is myoglobin,[30,31] made of about 150 residues *only*, while most interesting proteins are much larger. Typically, hexokinase has around 450 residues, tetrameric hemoglobin 600, dimeric citrate synthase 900, and a complete immunoglobulin antibody 1400. By assuming an average number of 30 degrees of freedom per residue—with making use of extended atoms, as often done—one finds that the total numbers N of degrees of freedom for the abovementioned systems are around 13,500, 18,000, 27,000, and 42,000, respectively. The reason why applications have been restricted so far to small proteins is as follows: the normal mode analysis method requires the diagonalization of $3N_a$ order matrices (N_a being the number of atoms) and most quick diagonalization methods for getting all the eigenvalues require the whole matrix to be loaded into the computer memory. Expressing the input matrix in Cartesian coordinates with double precision, this makes about 750 Mbytes of memory for hexokinase. Such a computer memory amount is larger than what is currently available to most computer-using scientists.

When matrix sparsity is taken into account, or when hard variables (bond lengths and valence angles) are omitted, less computer memory is required.[32-34] The possibilities to optimize such computer codes, however, are lower, so that the CPU (central processing unit) time, roughly proportional to N^3, now becomes the limiting step. Note that for *symmetrical* protein multimers, group theory can be put to profit in the diagonalization of such large matrices.[35] Protein multimers built from monomers with less than 1500 atoms are rare, however. On the other hand, more dedicated diagonalization methods, such as those used in quantum chemistry, usually require good starting approximations of the eigenvectors. In this spirit, an adapted version of the Lanczos algorithm has been applied to the hinge-bending motion in lyzozyme.[15] This system is particularly well suited to this method since a trial eigenvector can be inferred with great accuracy, the hinge-bending motion involving a small rotation of the two domains with respect to each other (about 3°).

Recently, efforts have been made to find methods requiring less computer memory or CPU time than the general ones, while making less assumptions on the eigenvectors than the dedicated ones. Two promising methods have been proposed, which share the following common ideas. First, they take profit from the idea that the structure of biological macromolecules can be considered as roughly linear. In proteins, the strongest interactions occur between neighbor components (i.e., amino acids or sets of amino acids), while interactions between distant components are relatively weak, indeed. Thus, when combined, low-frequency modes of single components are good—*but not good enough*—approximations of low-frequency normal modes of proteins.[36,37]

The two methods, however, differ in the way the first-step approximations are refined. In the first one, when computing the low-frequency modes of single components, the interactions between nearest neighbors are taken into account.[36] The matrix to be diagonalized is then expressed in the subspace of the low-frequency modes obtained in the first-step approximation, and the corresponding reduced eigenvalue problem is solved. Such a method, which is consistent with the Rayleigh–Ritz formulation of vibration problems treatment,[38] is supposed to only lead to upper bounds of the exact eigenvalues, but it happens to give very accurate results when applied to small test cases such as polypeptides. However, applications to large proteins have not yet been reported. In the second method, the matrix to be diagonalized is first expressed in a subset of local low-frequency modes and the corresponding reduced eigenvalue problem is solved.[37] Then, in order to reach the exact eigenvectors, an iterative approach is used. Each iteration is made of several steps, a subset of the eigenvector Cartesian coordinates being changed at each step, while the eigenvectors corresponding to the lowest eigenvalues found previously are retained. When applied to small protein cases, this method converged within a few iterations, without significant errors in the chosen low-frequency range. However, since diagonalizations of a rather large matrix are required at each iteration, this method is expected to be time-consuming.

In the present paper, a third method is proposed, which leads to exact results and is expected to be less CPU time-consuming than the two previous ones. It is a synthesis of both that is based on a perturbative approach currently used in quantum chemistry. Here, the first-step approximation for the low-frequency modes is not built through some frequency cutoff criterion, but by making use of a simple physical idea. A protein chain can be seen as being made of rigid components linked together. These components may be the amino acids or sets of amino acids such as those defining secondary structures or motifs. The combination of the translation and rotation motions of these rigid components can be expected to provide a reasonable approximate description of the lowest frequency modes. This philosophy is pictured in Figure 1. Studies on DNA have shown[39] that an approximation of this kind is very accurate for normal modes whose frequencies are lower than 380 cm^{-1}. In this work, the accuracy of such an approximation is tested on model proteins. The efficiency of the method will be particularly assessed with respect to the physical choice of the blocks of atoms remaining roughly rigid in the low-frequency modes.

METHOD

The normal modes of a system are determined by solving the following secular equations:

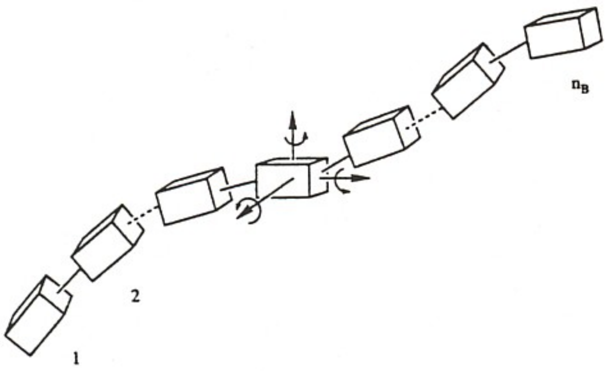

Figure 1. Principle of the low-resolution step. The macromolecule strand is divided into various blocks, whose local translations and rotations (arrows) provide the basis of a subspace in which the energy second-derivative matrix is expressed and diagonalized.

$$\begin{aligned} k_{11}x_1 + k_{12}x_2 + \cdots k_{1N}x_N &= \omega^2 m_1 x_1 \\ k_{21}x_1 + k_{22}x_2 + \cdots k_{2N}x_N &= \omega^2 m_2 x_2 \\ \cdots\cdots\cdots \\ k_{N1}x_1 + k_{N2}x_2 + \cdots k_{NN}x_N &= \omega^2 m_N x_N \end{aligned} \quad (1)$$

where the x_i are the Cartesian displacements of the atoms with respect to their equilibrium positions, $k_{ij} = (\partial^2 V)/(\partial x_i \partial x_j)$ are the second derivatives of the potential energy with respect to x_i, m_i are the atomic masses, ω is an eigen angular frequency, and N is the number of degrees of freedom of the system.[40] Equations (1) can also be written in matrix notation:

$$\boldsymbol{Kx} = \omega^2 \boldsymbol{Mx} \quad (2)$$

where

$$\boldsymbol{K} = \begin{bmatrix} k_{11} & k_{12} & \cdot & k_{1N} \\ k_{21} & k_{22} & \cdot & k_{2N} \\ \cdot & \cdot & \cdot & \cdot \\ k_{N1} & k_{N2} & \cdot & k_{NN} \end{bmatrix};$$

$$\boldsymbol{M} = \begin{bmatrix} m_1 & 0 & \cdot & 0 \\ 0 & m_2 & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \cdot & m_N \end{bmatrix}; \quad \boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \\ \cdot \\ x_N \end{bmatrix}$$

Multiplying both sides of Eq. (2) by $\boldsymbol{M}^{-(1/2)}$ leads to the transformed eigenvalue equation:

$$\boldsymbol{Hq} = \omega^2 \boldsymbol{q} \quad (3)$$

where $\boldsymbol{H}$ is the Hessian matrix and $\boldsymbol{q}$ the mass-weighted Cartesian coordinate vector:

$$\boldsymbol{H} = \boldsymbol{M}^{-(1/2)}\boldsymbol{K}\boldsymbol{M}^{-(1/2)}; \quad \boldsymbol{q} = \boldsymbol{M}^{1/2}\boldsymbol{x}$$

In the following, it will be convenient to use the more general Dirac notation,[41] which allows us to write the eigenvalue problem (3) in the form

$$H|q> = \omega^2|q> \quad (4)$$

where $|q>$ is an eigenvector of the Hessian operator H. To understand the relationship between notations (3) and (4), one can say that Eq. (3) is the matrix representation of Eq. (4) in the basis of the $|i>$ vectors arising from the mass-weighted Cartesian displacements of the atoms. With such a notation, the components of the Hessian matrix $\boldsymbol{H}$ in the basis of $|i>$ are denoted:

$$H_{ij} = \langle i|H|j\rangle = \frac{1}{\sqrt{m_i m_j}} k_{ij}$$

Similarly, the components of $|q>$ in this basis are

$$\langle i|q\rangle = \sqrt{m_i} x_i$$

In particular, the closure relation

$$\sum_{i=1}^{N} |i><i| = 1$$

expresses the completeness of the mass-weighted Cartesian basis set, while the relation

$$\sum_{i=1}^{N} |q_i><q_i| = 1$$

expresses the completeness of the basis set of the eigenvectors associated with the normal modes. In this work, however, we will not be interested by the determination of all normal modes of a macromolecule, but only by a few of them (the n lowest ones). This will be done through the prior determination of $6n_B$ (see below) approximate low-frequency eigenvectors denoted $|q_i^L>$, and $N - 6n_B$ approximate high-frequency eigenvectors denoted $|q_i^H>$, in such a way that these approximate eigenvectors also form a complete basis set:

$$\sum_{i=1}^{6n_B} |q_i^L><q_i^L| + \sum_{i=6n_B+1}^{N} |q_i^H><q_i^H| = 1$$

The representation of vectors and operators in such a basis happens to be intermediate between the initial Cartesian representation and the normal mode representation in which H is diagonal. Such a low-resolution description of the harmonic dynamics will be first presented. Next, we shall describe the derivation of exact solutions.

Low-Resolution Step

The macromolecule is first divided into n_B blocks. A block can be made of one or a few residues. The basic idea here is to assume that a good approximation of the n lowest frequency eigenvectors, those we are interested in, can be obtained by linear combination of vectors associated with the local displacements of the various blocks, namely their translations and rotations (see Figure 1). The approximate low-frequency eigenvectors $|q_i^L>$ are thus obtained by diagonalizing a matrix of order $6n_B$ representing H in the basis of the vectors arising from the six displacements of each of the n_B blocks (such modes are therefore delocalized over the entire molecule).

In a second step, the high-frequency eigenvectors $|q_i^H>$ localized in each block are determined in the following way. Let P_1 be the orthogonal projector associated with block I. It can be decomposed into a projector P_I^L associated with the translations and rotations of the block, and a projector P_I^H that is its orthogonal complement. The projectors associated with the corresponding subspaces are defined as

$$P_I = P_I^L + P_I^H \quad \text{with} \quad \sum_{I=1}^{n_B} P_I = 1$$

where 1 is the $N \times N$ identity operator. The diagonalization of the projected Hessian

$$H_I = P_I^H H P_I^H$$

provides $N_I - 6$ approximate high-frequency eigenvectors.

The set $\{|q_i^H>\}$ of all eigenvectors computed for each block spans a subspace that is the orthogonal complement of the subspace spanned by the approximate low-frequency eigenvectors. The exact eigenvectors will be obtained by perturbating these approximate eigenvectors by the high-frequency ones. For proteins, the perturbative derivation of the n lowest frequency exact modes is not obvious at all since one has to deal with a highly degenerated low-frequency spectrum, and one cannot be assured of a one-to-one correspondence between the n lowest frequency $|q_i^L>$ and the n lowest frequency exact solutions. Instead of considering a one-to-one correspondence between the unperturbed eigenmodes and the exact ones as done in standard perturbation theory, the subspace of exact solutions (target space) will be here derived from the subspace of the unperturbed eigensolutions (model space) within the theory of effective Hamiltonians,[42] which is briefly recalled below.

Effective Hessian

The theory of effective Hamiltonians is centered on the derivation of a n-dimensional effective Hamiltonian that provides n exact eigenenergies and eigenvectors that are the projections in the model space of the n lowest exact solutions.[42] In our classical vibrational problem, one may determine a n-

dimensional effective Hessian, the eigenvalues of which will be the n lowest exact eigenvalues ω^2 of the macromolecule. The associated unperturbed projector is

$$P_0 = \sum_{i=1}^{n} |q_i^L><q_i^L|$$

while the projector associated with the complementary space is

$$Q_0 = \sum_{i=n+1}^{6n_B} |q_i^L><q_i^L| + \sum_{i=6n_B+1}^{N} |q_i^H><q_i^H|;$$

$$P_0 + Q_0 = 1$$

Hereafter, the Hessian operator H will be split into two parts: an unperturbed part H_0 and a perturbation V. The unperturbed part is defined as

$$H_0 = \sum_{i=1}^{6n_B} (\omega_i^L)^2 |q_i^L><q_i^L| + \sum_{i=6n_B+1}^{N} (\omega_i^H)^2 |q_i^H><q_i^H|$$

where $(\omega_i^L)^2 = \langle q_i^L|H|q_i^L\rangle$ and $(\omega_i^H)^2 = \langle q_i^H|H|q_i^H\rangle$. With the above definition, the matrix representation of H_0 in the basis set $\{|q_i^L>, |q_i^H>\}$ is diagonal. The effective Hessian can thus be written as

$$H^{\text{eff}} = P_0 H P_0 + P_0 V X$$

where $X \equiv Q_0 X P_0$ is the reduced wave operator that couples the model space to the complementary space. The diagonalization of H^{eff} provides the n lowest ω^2:

$$H^{\text{eff}} = \sum_{i=1}^{n} \omega_i^2 |q_i^0><q^{i,0}| \qquad (5)$$

where $|q_i^0> = P_0|q_i>$ is the projection in the model space of the exact eigenvectors $|q_i>$ and $|q^{i,0}>$ is the corresponding biorthogonal vector characterized by $\langle q^{i,0}|q_j^0\rangle = \delta_{ij}$. In this context, the problem of finding the n lowest frequency eigenmodes reduces to the determination of X, which obeys a Bloch-like equation of the form

$$X = \sum_{i=1}^{n} \frac{Q_0}{\omega_i^2 - H_0} V(1 + X)P_i \qquad (6)$$

where $P_i = |q_i^0><q^{i,0}|$ is the nonorthogonal projector associated with the eigen solutions of H^{eff}. From Eq. (6), Eq. (5) can be transformed into the perturbative expression:

$$H^{\text{eff}} = P_0 H P_0 + \sum_{i=1}^{n} P_0 V \frac{Q_0}{\omega_i^2 - H_0} V(1 + X)P_i \quad (7)$$

As a consequence, the determination of the reduced wave operator X can be obtained by solving Eq. (6) using the perturbation-iteration scheme described in Ref. 43.

Status with Respect to Other Procedures

Our method presents similarities and differences with respect to the two recent works mentioned above.[36,37] In these works, the macromolecule is also divided into smaller components from which local modes are extracted. These components are analogous to our blocks but the way the local information is used is different in the three approaches.

In Mouawad and Perahia's approach,[37] the local modes are combined linearly to get approximate low-frequency modes that are delocalized over the entire macromolecule. They are next improved iteratively by mixing with Cartesian coordinates until convergence. The procedure is very stable and converges rapidly, but the computational effort remains important since large matrices have to be diagonalized at each iteration. The method has been applied to small proteins but the authors did not report any computational time estimation as a function of N.

In Hao and Harvey's method,[36] the determination of local modes is improved from interactions with the neighboring blocks. The authors ground their method on a previous work by Ookuma and Nagamatus[44] but it can be formulated within the theory of effective Hamiltonians as well. Typically, expressions (16) and (17) in Ref. 36 are nothing else than effective Hessians, analogous to our expression (7). However, their effective interactions and effective Hessians are not used for the same purpose as ours. They are used to improve local modes while in our approach the effective Hessian (7) is associated with approximate low-frequency modes that are *delocalized* over the entire macromolecule and that are interacting with approximate high-frequency modes *localized* in the blocks. Moreover, in that method, several approximations are further incorporated such as the neglect of high-frequency modes or the use of "static conditions," which makes their formalism approximate while ours leads to exact solutions. From a computational point of view, their use of local effective Hessians no doubt reduces the dimensionality of the eigenvalue problem. However, the authors neither report calculations on real proteins (only polypeptides and a small DNA oligomer were considered) nor mention the dependence of the CPU time on the size of the problem.

When dealing with larger systems, these two above methods seem to be promising since they require reasonable computer memory, but they still have to prove their efficiency regarding the CPU time requirement. In our case, we show below how our method is nearly N^2 dependent on the size of the system.

CPU Time Requirement

Implementation of our method was first done to check whether it could work for proteins, especially regarding the convergence problems. We are currently optimizing the codes regarding the memory and CPU time requirements, so that significant quantitative information about computational performances is therefore not yet available. It is possible, however, to make some quantitative estimations of the N dependence. The initial low-resolution description requires the calculation of the elements of the Hessian matrix of dimension $6n_B$. This involves a number of multiplications proportional to N^2. On the other hand, the computational time for diagonalizing this matrix is proportional to $(6n_B)^3 = 216n_B^3$, while the computational time for getting the zero-order high-frequency modes in the various blocks is roughly proportional to $n_B(N/n_B)^3 = N^3/n_B^2$ (if all blocks are assumed to have the same dimension). The total computational time for diagonalizing all these matrices is therefore proportional to $216n_B^3 + N^3/n_B^2$. The optimal number of blocks can be obtained by minimizing this expression with respect to n_B. An immediate derivation yields to $n_B = 0.3N^{3/5}$ with a corresponding CPU time therefore proportional to $17N^{9/5}$. This power 9/5 happens to be slightly under the power 2 occurring in the calculation of the elements of the low-frequency Hessian matrix. Since the final perturbation-iteration step is also proportional to N^2, the overall numerical procedure appears to be rather well balanced, and should keep a N^2 dependence for very large proteins.

TEST CALCULATIONS

Three systems of increasing size and structural complexity have been tested: two regular α-helix oligomers, decaalanine and icosaleucin, and a real small protein, crambin. Decaalanine, $(\text{Ala})_{10}$, is taken with methyl groups at both ends, and icosaleucin, $(\text{Leu})_{20}$, is taken in its zwitterionic form. Starting from α-helix configurations, these two oligopeptides were optimized with Powell's algorithm within the standard CHARMM-19 force field and using extended atoms for CH, CH_2, and CH_3.[45] For decaalanine, all atom–atom interactions were taken into account. For icosaleucin, a cutoff at 7.5 Å was used in the calculation of nonbonded interactions, in conjunction with a switching function between 6.5 and 7.5 Å, and a shifting function for the electrostatic interactions.[45] The third system, crambin, is a globular protein of 46 amino acids, whose tridimensional structure is known to a high accuracy.[46] Although one of the smallest natural proteins, crambin exhibits both kinds of secondary structures (two α-helices and a β-sheet) and disulfide bridges, which makes it an ideal target for our methodological tests. This system was optimized with the same protocol as icosaleucin. As in a previous study, the optimized structure remains close to the x-ray geometry used as starting point.[26] In all three cases, the mass-weighted second-derivative matrix was calculated from two-points finite differences. Using these protocols, the lacunarities of the matrices for decaalanine, icosaleucin, and crambin are of 0, 34, and 81%, respectively.

Let us examine, first, how good our normal modes are at the end of the low-resolution step. This can be measured both from the value of the frequencies and from the quality of the corresponding vectors. To visualize the components of each vector, one may plot the module of atom displacements along the chain. This provides a convenient fingerprint for each mode and roughly indicates the location of the main deformations in the molecule. We also tried to use the modules of the displacements of the centers of gravity for each residue. This may help reduce the information for very large proteins, but in the present cases it did not bring much more help than the atom-by-atom plotting.

We will begin with a partitioning of the protein chains into as many blocks as residues. In other words, in this first *standard* partitioning, there is one residue per block. This makes 12 blocks for $(\text{Ala})_{10}$ (the two methylated ends being counted apart), 20 blocks for $(\text{Leu})_{20}$, and 46 blocks for crambin. Next, we have increased the size of the blocks to 2, 3, and more residues, thus dividing the total number of blocks by these corresponding factors.

For the standard one-residue-per-block division, the low-resolution frequencies and their evolution at each iteration of the perturbation-iteration process are plotted for the ten lowest modes of decaalanine in Figure 2 and for the five lowest modes of icosaleucin and crambin in Figures 3 and 4 respectively. In these curves, the starting values of the frequencies are sufficiently well located so that convergence to a reasonable accuracy—say four significant figures on the frequencies—is reached within a few iterations, namely less than 5 iterations for $(\mathrm{Ala})_{10}$ and less than 10 iterations for $(\mathrm{Leu})_{20}$ and crambin. Note, however, that the higher the frequency mode, the less accurate the starting point. This could make more difficult the obtention of higher frequency modes and further improvements could then have to be introduced in the procedure, such as the use of an intermediate space in the perturbation-iteration step, the diagonalization of a subspace larger than $6n_B$ in the low-resolution step, or the introduction of further degrees of hierarchy.

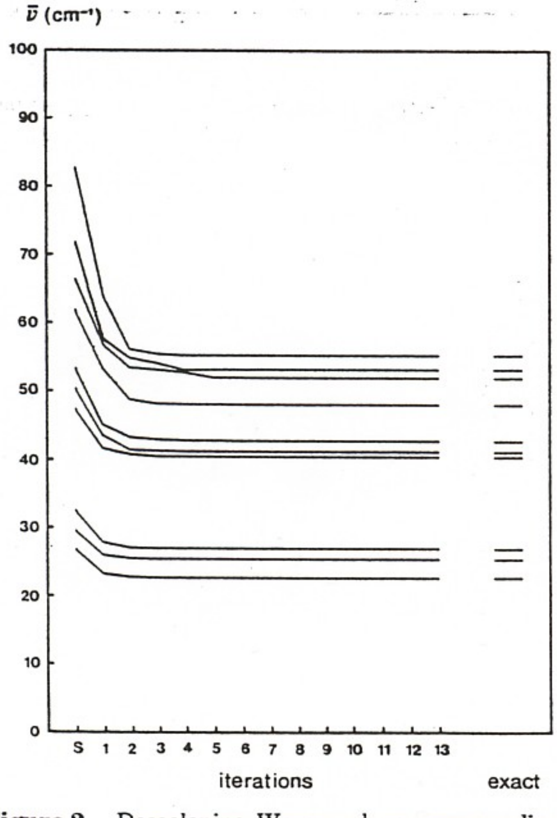

Figure 2. Decaalanine. Wavenumbers corresponding to the ten lowest normal modes in the standard (one-residue-per-block) partitioning as a function of the number of iterations in the perturbation process. The starting values, at left, result from the low-resolution step. The exact values are indicated at right.

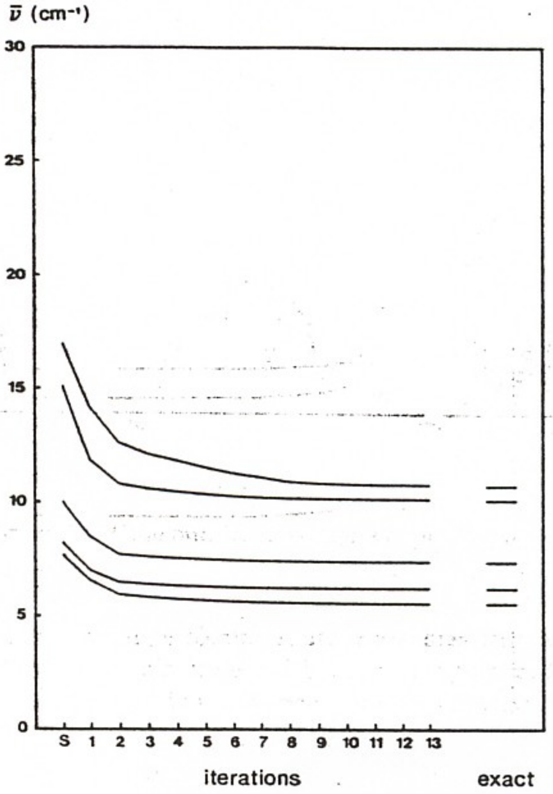

Figure 3. Icosaleucin. Wavenumbers corresponding to the five lowest normal modes in the standard partitioning.

In Figures 2 and 4, neat avoided crossings can be noticed early in the iteration process. This means that although our low-resolution modes are good approximations for obtaining the exact vectors, the starting ordering may be different from the final ordering. This is particularly striking in Figure 4 where the second low-resolution vector clearly relates to the third exact solution. We shall see below that these vectors are quite similar, indeed.

When the chain of blocks is less resolved, i.e., if we build our local translations and rotations over blocks that are bigger and less numerous, the starting vectors are not so good, but still accurate enough to yield the exact solutions within a few iterations. This is illustrated in Figure 5 for decaalanine, here divided into six blocks, four of which including two residues. See how the starting frequencies, at left, are higher than in Figure 2.

Let us now have a look at the quality of the low-

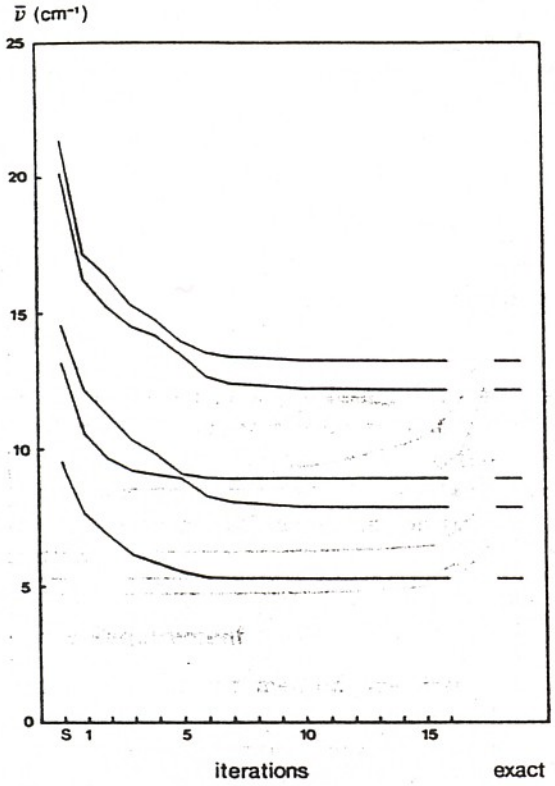

Figure 4. Crambin. Wavenumbers corresponding to the five lowest normal modes in the standard partitioning.

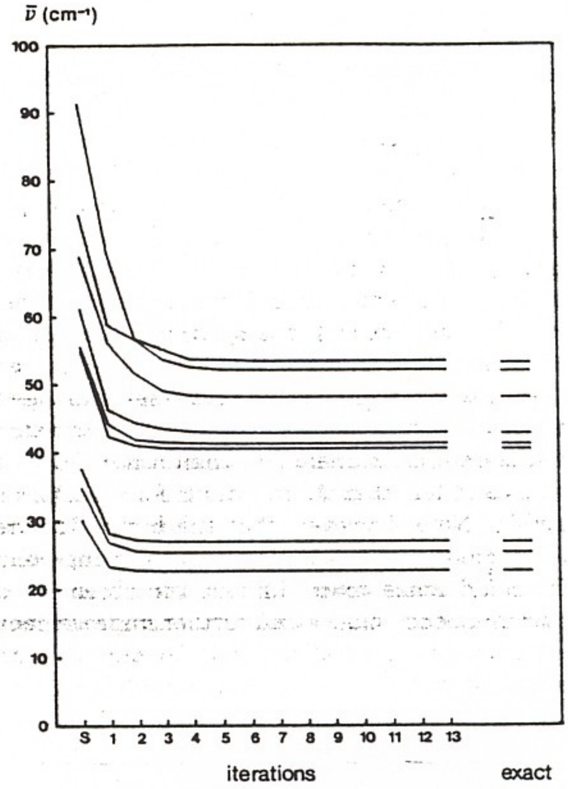

Figure 5. Decaalanine. Obtention of the wavenumbers corresponding to the nine lowest normal modes in the two-residue-per-block partitioning.

resolution modes regarding their components over the modules of the atom displacements. For the standard partitioning into one residue per block, the lowest vector resulting from the diagonalization in the subspace of local translations and rotations is always a very good approximation for the lowest mode, as exemplified for decaalanine and crambin in Figures 6 and 7, respectively. For the next modes, the quality is slightly reduced but remains quite reasonable, as exemplified for the third lower modes of decaalanin and crambin in Figures 8 and 9, respectively. In the latter case, it is the second low-resolution mode that generates the third exact mode, because of the avoided crossing visible in Figure 4. For the sixth lower mode of decaalanine, the approximation provided by the low-resolution step is now less conspicuous (Figure 10, top). The general shape of the mode, however, remarkably gets into position at as early as iteration one of the perturbation process (Figure 10, bottom).

As the number of blocks is reduced, one may expect a degradation of our zero-order low-resolution modes. Actually, most vectors happen to resist fairly well to this treatment, the most robust being of course the lowest one. For decaalanine, this mode is plotted in Figure 11 as it appears at the end of the low-resolution step, for both a 12-block or a 3-block partitioning of the helix chain. Both cases ex-

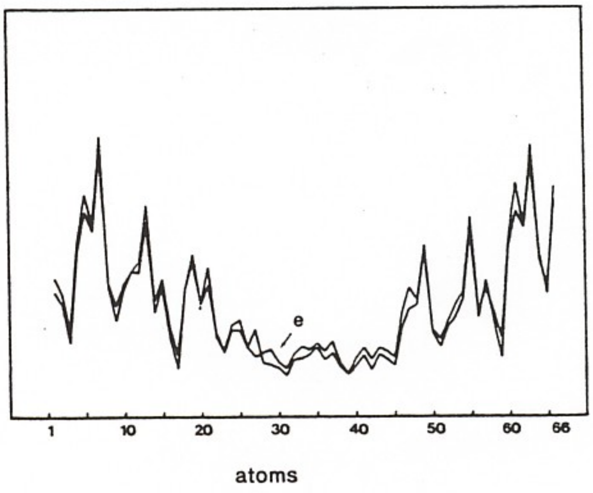

Figure 6. Decaalanine. Composition of the lowest normal mode expressed by the displacement modules for each atom along the protein chain. The curve labeled *e* corresponds to the exact eigenvector. The other curve results from the low-resolution step in the standard partitioning.

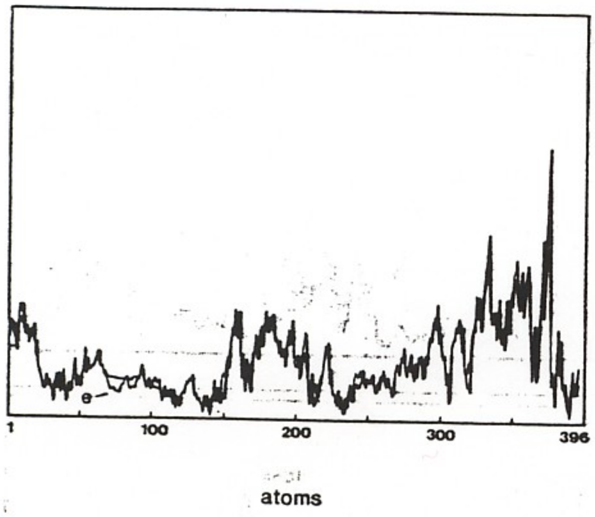

Figure 7. Crambin. Composition of the lowest normal mode expressed by the displacement modules of each atom. Comparison of the exact solution *e* with the starting low-resolution eigenvector in the standard partitioning.

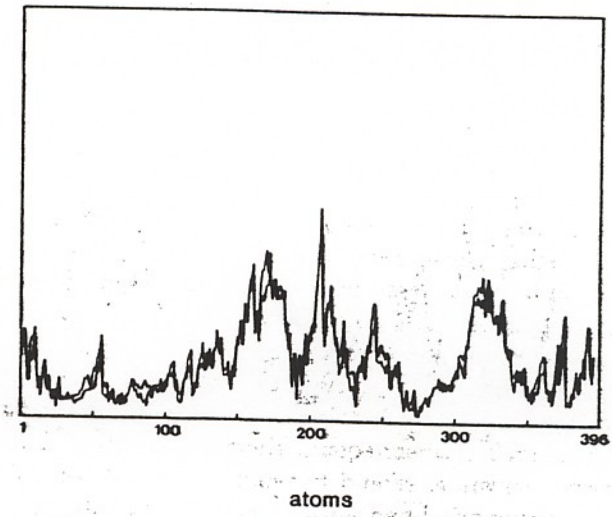

Figure 9. Crambin. Same as Figure 7, for the third lowest mode.

hibit similar shapes. For the third lower mode of decaalanine, the same comparison makes clear this mode is less well approximated with a 3-blocks-only partitioning (Figure 12). For crambin, when the 46-residues chain is divided into 10 blocks of five residues each on average, the low-resolution step is still a good starting point for both the lowest mode and the third lower mode (Figure 13).

We next tried to individualize a secondary structure in crambin, thus incorporating into the method some information about its *known* three-dimensional structure. For doing so, we have clusterized the eight residues of the small α-helix (residues 23–30) into a single block, all other residues still contributing one block each. This does not perturb too much the starting low-resolution description, as

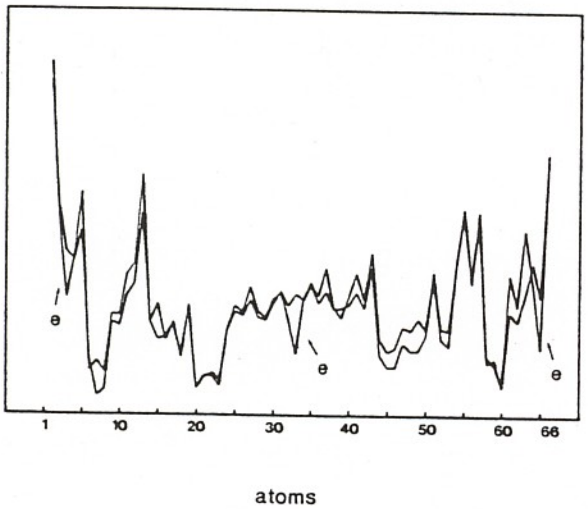

Figure 8. Decaalanine. Same as Figure 6, for the third lowest mode.

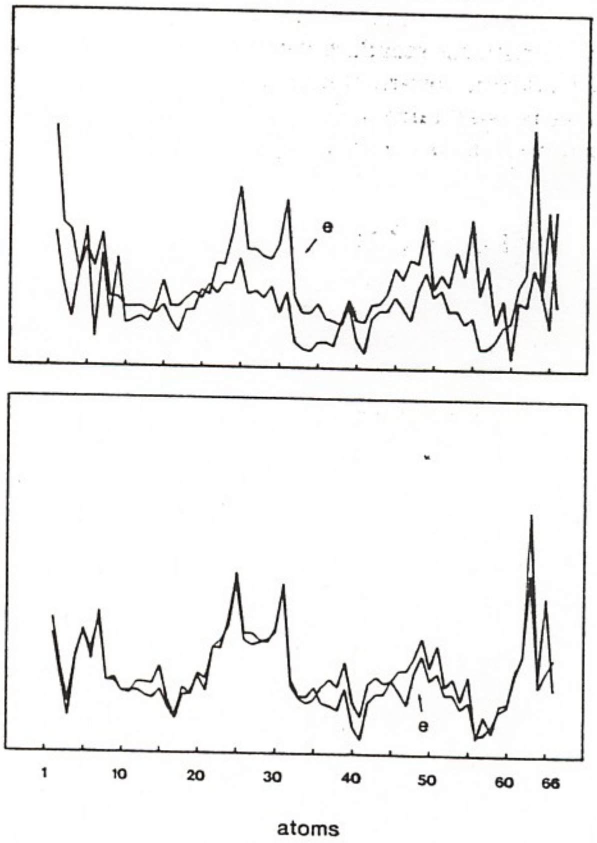

Figure 10. Sixth lowest mode of decaalanine in the standard partitioning. Exact vs low-resolution step eigenvector (top), and how it gets into position at the first iteration of the perturbation process (bottom).

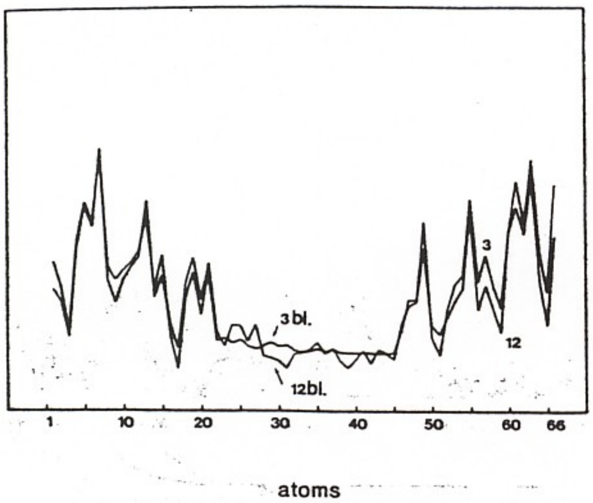

Figure 11. Lowest mode of decaalanine. Comparison of the low-resolution step for partitions of the chain into 12 blocks and 3 blocks.

shown for the two lowest modes in Figure 14. Going a step further, the large α-helix (residues 6–20) is further contracted into two blocks of 7 and 8 residues, leading to a set of 26 blocks only. When compared with the standard partitioning into 46 blocks, the starting low-resolution modes again appear to have resisted fairly well to this secondary-structure contraction, as can be seen in Figure 15.

TOWARD LARGER SYSTEMS

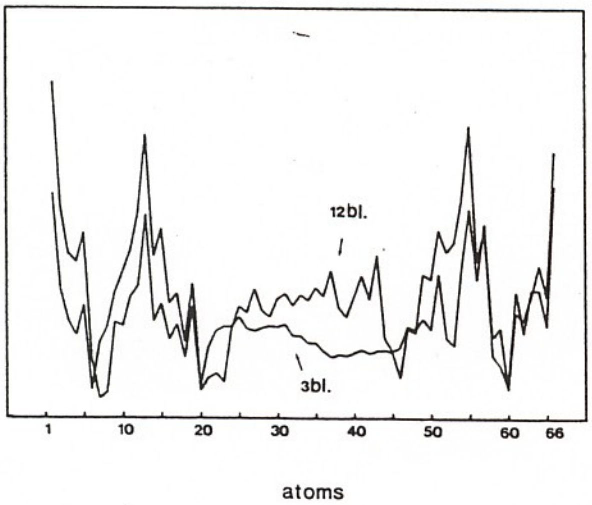

Figure 12. Decaalanine. Same as Figure 11 for the third lowest mode.

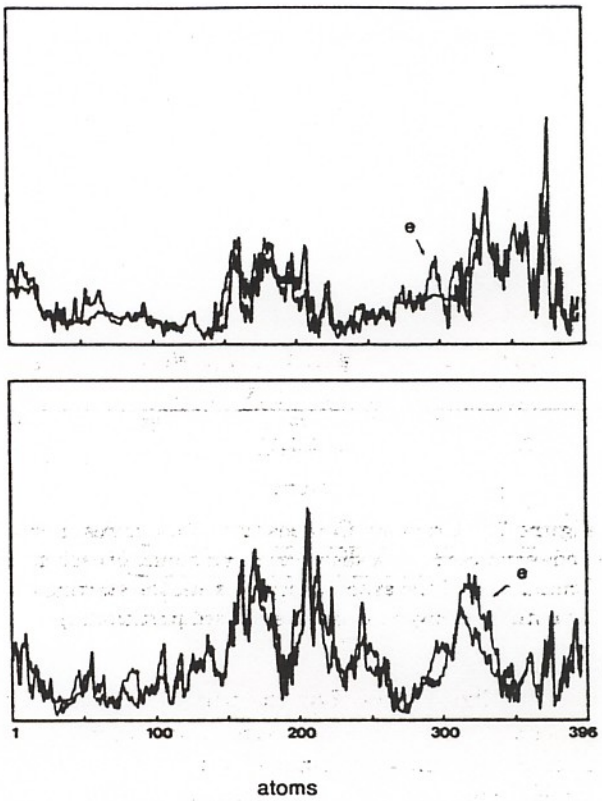

Figure 13. Crambin, first (top) and third (bottom) lowest modes. Comparison of the exact solution with the low-resolution step for the five-residue-per-block partitioning.

Considering proteins or polypeptides as being made up of rigid components linked together have yielded good zero-order approximations for the lowest frequency normal modes, especially regarding the eigenvectors. Nevertheless, although usually correctly ordered at the low-resolution step, the corresponding frequencies may differ from their exact values by as much as a factor of two (remember Figure 4). So, any further improvement of these starting frequencies would be welcome when we shall address much larger systems. In the present tests, a one-residue-per-block partitioning was always found to be efficient, but a strategy of block building that would take into account any information about secondary or tertiary structures would no doubt improve these low-resolution frequencies. This is all the more true that in such cases motions of rigid structural elements should be more meaningful than in our small systems herein tested.

In this paper, only small systems have been investigated but we now intend to address large proteins. In Table I are summarized the data and constraints characterizing such systems up to 1500 residues. This size—typically that of a complete antibody molecule—can be considered as an upper

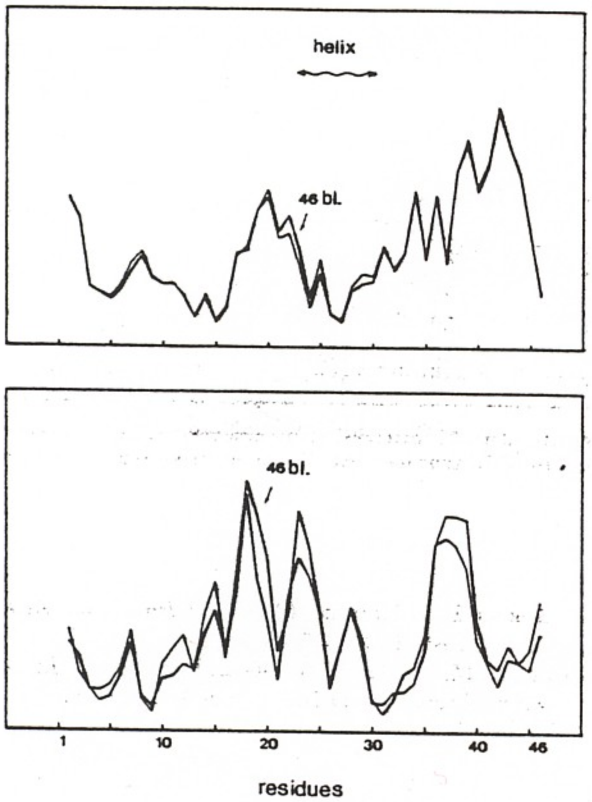

Figure 14. Crambin, low-resolution modes. Comparison of the standard partitioning with one in which the small helix is clustered into a single block. Top: lowest mode; bottom: second lowest mode. The eigenvectors are here expressed in the displacement modules of the center of mass for each residue.

limit to be handled, but a lot of interesting functional proteins range between 100 and 1000 residues (corresponding to molecular weights of 10–100 kD). The specifications listed in Table I are obtained from a partition optimizing the CPU time, as explicated at the end of the Method section. The number of blocks and their optimal size smoothly increases with N in the range of current proteins. Even in extreme cases (1500 residues), the corresponding elementary computational steps remains reasonable. The block size always remains small (< 240) while the dimension of the $6n_B$ low-resolution Hessian matrix is at worst of order 1128. Such a matrix can now be diagonalized within a few minutes on currently available workstations. Concerning the disc storage requirement of the input mass-weighted Hessian matrix, for the range 500–1000 residues, one needs 300–1000 Mbytes, which corresponds to disk capacities currently offered on workstations.

The final perturbation step of our method requires a computational effort proportional to nN^2 with n being the number of searched low-frequency modes. The procedure therefore exhibits a consistent roughly N^2 dependence in each step. Let us recall that full diagonalization of the complete Hessian matrix has a CPU time proportional to N^3. Thus, as a function of N, the estimate of computational time required by our method seems quite reasonable and should be suitable for investigating large systems corresponding to very large values of N ($> 10{,}000$). Beyond these capabilities, the proposed method presents additional advantages. First, the low-resolution step can be fully parallelized since the determination of the high-frequency modes in each block can be made independently. Table I indicates that the number of blocks is always lower than 200, which is, on another hand, the number of processors becoming now available on most efficient parallel machines. Our approach has a last advantage if we want to refine or improve it: it is entirely formulated within the framework of the theory of

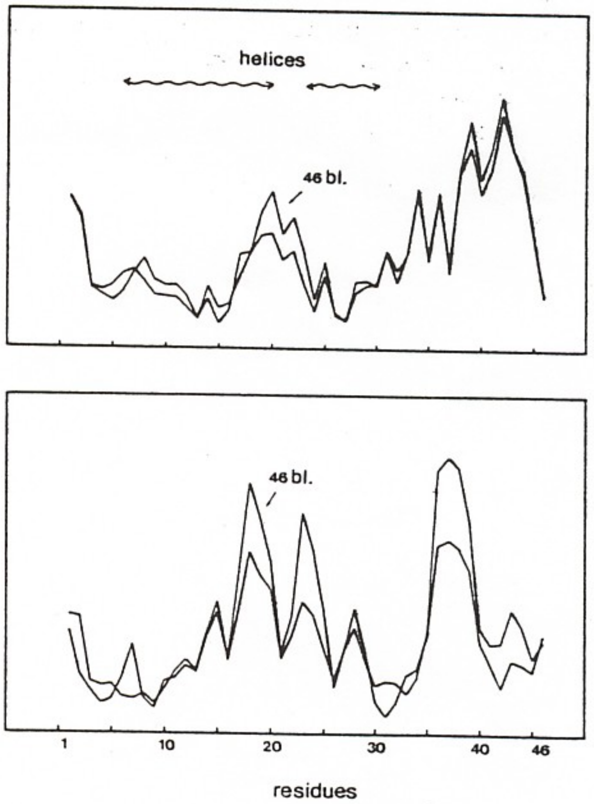

Figure 15. Crambin, low-resolution modes. Comparison of the standard partitioning (46 blocks) with a 26-block partitioning corresponding to the contraction of the small helix into a single block, and that of the large helix into two blocks. Top: lowest mode; bottom: second lowest mode. The eigenvectors are expressed in the displacement modules of the center of mass for each residue.

Table I Summary of Mean Data for Proteins of Different Sizes[a]

Number of Residues	Molecular Weight (kD)	Size of the Matrix N	Number of Blocks n_B	Residues per Block	Block Size	Size of the $6n_B$ Matrix	Minimal Disk Storage Requirement for the Input Matrix (Mbytes)
10	1	300	10	1	30	60	< 1
50	5	1500	25	2	60	150	3
100	10	3000	33	3	90	540	11
500	55	15,000	100	5	150	600	270
1000	110	30,000	167	6	180	1002	1080
1500	165	45,000	188	8	240	1128	2430

[a] The average parameters for a mean residue are as follows. Degrees of freedom: 30 (extended atoms assumed); molecular weight: 110. The number of blocks is taken in the optimal partition that minimizes the CPU time (see text). The input mass-weighted second-derivative matrix is assumed to have a lacunarity of 80%.

the effective Hamiltonians, and we can take benefit of the large experience accumulated over years in this field. In particular, significant progress in the theory and efficiency in the algorithm could be attained by constructing effective Hessians associated with various structural hierarchies such as domains, motifs and secondary structures.[47] Work is in progress along these lines.

We wish to thank Jiri Savrda and Gabin Treboux for their assistance in the code writing. We also thank Lilianne Mouawad and David Perahia for fruitful discussions and for providing us with their manuscript prior to its publication.

REFERENCES

1. Bennett, W. S. & Steitz, T. A. (1980) *J. Mol. Biol.* **140,** 183–209.
2. Bennett, W. S. & Steitz, T. A. (1980) *J. Mol. Biol.* **140,** 210–230.
3. Remington, S., Weigand, G. & Huber, R. (1982) *J. Mol. Biol.* **158,** 111–152.
4. Wiegand, G. (1984) *J. Mol. Biol.* **174,** 205–219.
5. Fermi, G. (1975) *J. Mol. Biol.* **97,** 237–256.
6. Shaanan, B. (1983) *J. Mol. Biol.* **171,** 31–59.
7. Faber, H. R. & Matthews, B. W. (1990) *Nature* **348,** 263–266.
8. Lesk, A. M. & Chothia, C. (1984) *J. Mol. Biol.* **174,** 175–191.
9. Wiegand, G. & Remington, S. (1986) *Ann. Rev. Biophys. Biophys. Chem.* **15,** 97–117.
10. McCammon, J. A., Gelin, B. R., Karplus, M. & Wolynes, P. G. (1976) *Nature* **262,** 325–326.
11. Mao, B., Pear, M. R., McCammon, J. A. & Quiocho, F. A. (1982) *J. Biol. Chem.* **257,** 1131–1133.
12. Go, N., Noguti, T. & Nishikawa, T. (1983) *Proc. Natl. Acad. Sci. USA* **80,** 3696–3700.
13. Brooks, B. & Karplus, M. (1983) *Proc. Natl. Acad. Sci. USA* **80,** 6571–6575.
14. Levitt, M., Sander, C. & Stern, P. S. (1983) *Int. J. Quant. Chem. Quant. Biol. Symp.* **10,** 181–199.
15. Brooks, B. & Karplus, M. (1985) *Proc. Natl. Acad. Sci. USA* **82,** 4995–4999.
16. Gibrat, J. F. & Go, N. (1990) *Proteins* **8,** 258–279.
17. Horiuchi, T. & Go, N. (1991) *Proteins* **10,** 106–116.
18. Swaminathan, S., Ichiye, T., Van Gunsteren, W. F. & Karplus, M. (1982) *Biochemistry* **21,** 5230–5241.
19. Levy, R. M., Perahia, D. & Karplus, M. (1982) *Proc. Natl. Acad. Sci. USA* **79,** 1346–1350.
20. Hoffman, D. K., Nord, R. S. & Ruedenberg, K. (1986) *Theor. Chim. Acta* **69,** 265–279.
21. Smith, C. M. (1990) *Int. J. Quant. Chem.* **37,** 773–783.
22. Cerjan, C. J. & Miller, W. H. (1981) *J. Chem. Phys.* **75,** 2800–2806.
23. Ech-Cherif El-Kettani, M. A. & Durup, J. (1992) *Biopolymers* **32,** 561–574.
24. Karplus, M. & Kushick, J. (1981) *Macromolecules* **14,** 325–332.
25. Levy, R. M., Karplus, M., Kushick, J. & Perahia, D. (1984) *Macromolecules* **17,** 1370–1374.
26. Teeter, M. M. & Case, D. A. (1990) *J. Phys. Chem.* **94,** 8091–8097.
27. Noguti, T. & Go, N. (1985) *Biopolymers* **24,** 527–546.
28. Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. & Teller, E. (1953) *J. Chem. Phys.* **21,** 1087–1092.
29. Diamond, R. (1990) *Acta Cryst. A* **46,** 425–435.
30. Seno, Y. & Go, N. (1992) *J. Mol. Biol.* **216,** 95–109.
31. Seno, Y. & Go, N. (1992) *J. Mol. Biol.* **216,** 111–126.
32. Noguti, T. & Go, N. (1982) *Nature* **296,** 776–778; Noguti, T. & Go, N. (1983) *J. Phys. Soc. Jpn.* **52,** 3285–3288.
33. Levitt, M., Sander, C. & Stern, P. S. (1985) *J. Mol. Biol.* **181,** 423–447.
34. Durup, J. (1991) *J. Phys. Chem.* **95,** 1817–1829.

35. Simonson, T. & Perahia, D. (1992) *Biophys. J.* **61,** 410–427.
36. Hao, M.-H. & Harvey, S. C. (1992) *Biopolymers* **32,** 1393–1405.
37. Mouawad, L. & Perahia, D. (1993) *Biopolymers* **33,** 599–611.
38. Oden, J. T. & Reddy, J. N. (1983) *Variational Methods in Theoretical Mechanics,* Springer Verlag, New York.
39. Saxena, V. K., Van Zandt, L. L. & Schroll, W. K. (1989) *Chem. Phys. Lett.* **164,** 82–86.
40. Goldstein (1950) *Classical Mechanics,* Addison-Wesley, Reading, MA.
41. Dirac, P. A. M. (1958) *The Principles of Quantum Mechanics,* Clarendon Press, Oxford.
42. Durand, Ph. & Malrieu, J.-P. (1987) in *Ab-Initio Methods in Quantum Chemistry—I.,* Lawley, K. P., Ed., John Wiley & Sons, New York, pp. 352–412.
43. Durand, Ph. (1983) *Phys. Rev. A.* **28,** 3184–3192.
44. Ookuma, M. & Nagamatus, A. (1984) *Bull. Jpn. Soc. Mech. Engin.* **27,** 529–533, 1288–1293.
45. Brooks, B. M., Bruccoleri, R. E., Olafson, B. D., States, D. J., Swaminathan, S. & Karplus, M. (1983) *J. Comput. Chem.* **4,** 187–217.
46. Teeter, M. M. (1984) *Proc. Natl. Acad. Sci. USA* **81,** 6014–6018.
47. Branden, C. & Toose, J. (1991) *Introduction to Protein Structure,* Garland Publications, New York.

Received July 13, 1993
Accepted September 8, 1993