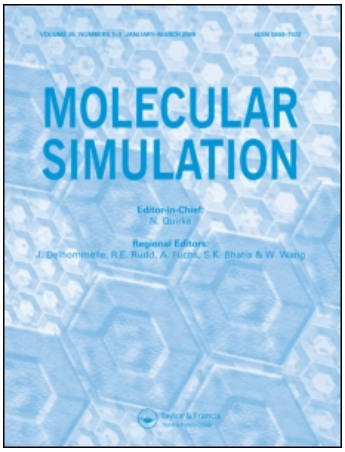

Molecular Simulation
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713644482

Projection Methods for the Analysis of Complex Motions in Macromolecules

Konrad Hinsen[a]; Gerald R. Kneller[a]
[a] Centre de Biophysique Moléculaire (CNRS UPR 4301), Orléans Cedex 2, France

To cite this Article Hinsen, Konrad and Kneller, Gerald R.(2000) 'Projection Methods for the Analysis of Complex Motions in Macromolecules', Molecular Simulation, 23: 4, 275 — 292
To link to this Article: DOI: 10.1080/08927020008025373
URL: http://dx.doi.org/10.1080/08927020008025373

Molecular Simulation, 2000, Vol. 23, pp. 275–292
Reprints available directly from the publisher
Photocopying permitted by license only

PROJECTION METHODS FOR THE ANALYSIS OF COMPLEX MOTIONS IN MACROMOLECULES

KONRAD HINSEN and GERALD R. KNELLER*

Centre de Biophysique Moléculaire (CNRS UPR 4301), Rue Charles Sadron, 45071 Orléans Cedex 2, France

(Received June 1999; accepted July 1999)

In studies of macromolecular dynamics it is often desirable to analyze complex motions in terms of a small number of coordinates. Only for simple types of motion, *e.g.*, rigid-body motions, these coordinates can be easily constructed from the Cartesian atomic coordinates. This article presents an approach that is applicable to infinitesimal or approximately infinitesimal motions, *e.g.*, Cartesian velocities, normal modes, or atomic fluctuations. The basic idea is to characterize the subspace of interesting motions by a set of (possibly linearly dependent) vectors describing elementary displacements, and then project the dynamics onto this subspace. Often the elementary displacements can be found by physical intuition. The restriction to small displacements facilitates the study of complicated coupled motions and permits the construction of collective-motion subspaces that do not correspond to any set of generalized coordinates.

As an example for this technique, we analyze the low-frequency normal modes of proteins up to $\approx 20\,\mathrm{THz}$ ($600\,\mathrm{cm}^{-1}$) in order to see what kinds of motions occupy which frequency range. This kind of analysis is useful for the interpretation of spectroscopic measurements on proteins, *e.g.*, inelastic neutron scattering experiments.

Keywords: Protein dynamics; normal modes; trajectory analysis

1. INTRODUCTION

Numerical methods such as Molecular Dynamics simulations or normal mode analysis have become standard tools for studying the dynamics of macromolecules. Once they have been verified by calculating experimentally observable quantities and comparing to experiment, the results can be

*Corresponding author. e-mail: kneller@cnrs-orleans.fr

analyzed further in order to obtain quantities that are not readily accessible to experiment and thus reach a better understanding of the dynamical processes in the system. A field in which the combination of experiment and simulation has been particularly successful is inelastic neutron scattering on proteins, where simulation is essential to interpret experimental spectra [1, 2].

Numerical calculations are usually performed using Cartesian coordinates for the individual atoms in the system. However, many quantities of interest are defined in terms of larger subgroups of a macromolecule, such as amino acid residues in proteins. In some cases a rigid-body analysis yielding the global rotation and translation of molecular subgroups is useful [3–6]. To be able to describe more complicated motions, one might want to use an arbitrary set of generalized coordinates. However, this approach often requires complicated calculations, in particular for finding the derivatives of the generalized coordinates, and this work has to be done from scratch for each new coordinate. An example is normal mode analysis in torsional angle space by explicit coordinate transformation [7].

A simpler analysis is possible if infinitesimal or at least bounded motions of small amplitude are considered. This allows to work with arbitrary infinitesimal *displacements*, which need not be differentials of coordinates. Examples for infinitesimal motions in molecular simulations are velocities and normal mode vectors, but also any bounded motion of sufficiently small amplitude that can to a good approximation be considered infinitesimal, *e.g.*, fluctuations around the native state in proteins. For some kinds of motion, *e.g.*, rotations, infinitesimal motions are significantly easier to handle than finite ones. In this article we develop an approach for analyzing infinitesimal motions or displacements by means of projection methods. An advantage of our method is the ease of combining several elementary displacements. For most practically important cases, combinations of a few standard types are sufficient, such that no significant effort is required to apply the technique to a new situation.

As an application, we present an analysis of the low-frequency normal modes of proteins. Like in all physical systems, low frequencies correspond to collective motions, whereas high frequencies describe localized motions. At the upper end of the frequency spectrum, around 100 THz ($3000\,cm^{-1}$), are the bond-stretching vibrations involving hydrogen atoms. Moving towards lower frequencies, there are the bond stretching vibrations between two heavy atoms, bond angle vibrations, rigid-body motions of larger chemical groups, internal deformations of residues, residue rigid-body motions, secondary structure deformations, and finally large scale collective motions, such as domain motions. The high-frequency part of the spectrum has been

analyzed in detail by classical spectroscopy techniques on small peptide chains *e.g.* [8]. The very low frequency motions have been studied in detail as well, because they contain the highly specific domain motions which determine a protein's function [9]. However, these motions occupy only a tiny part of the frequency spectrum; normal mode calculations of proteins of various size show that the number of modes describing domain motions is roughly equal to a hundredth of the number of residues [10].

The large frequency interval between domain motions and single-residue vibrations is much less well understood. The only study of motions in a particular low-frequency interval that we are aware of is an analysis of diffusive motion up to a time scale of ≈ 100 ps, which showed that the major contribution comes from liquid-like rigid-body motion of the sidechains [5, 6]. This result was obtained by eliminating the sidechain deformations from a molecular dynamics trajectory using rigid-body fits and comparing the spectrum of this modified trajectory to the original one. In this paper, we analyze the low-frequency vibrational motions using normal mode techniques. Normal mode frequency spectra can be related to neutron scattering spectra [11, 12], and in spite of the inherent approximations (no anharmonic or solvent effects) they have provided much useful information about protein dynamics. By projecting each normal mode on subspaces that contain well-defined motions, we show which kinds of motion contribute to specific parts of the spectrum.

2. DECOMPOSITION OF INFINITESIMAL MOTIONS

Infinitesimal motions describe directions in the configuration space of a physical system. As already mentioned, the most important examples for molecular simulations are velocities describing the directions of a trajectory at a given time, and normal mode vectors describing the directions along which an harmonic system can oscillate at a single frequency. However, many finite but bounded motions can be considered infinitesimal to a good approximation. For example, it is common to analyze protein trajectories obtained from Molecular Dynamics trajectories by Principal Component Analysis of the atomic fluctuation matrix [13], which yields a set of directions in configuration space onto which trajectories are projected as if they were infinitesimal motions. In the following, a given infinitesimal motion is described by a vector $\mathbf{v}_i$ for each atom i. In a system consisting of N atoms, there are thus in total $3N$ infinitesimal coordinates, which together form a vector in the $3N$-dimensional configuration space.

The basic idea of our analysis is the projection of infinitesimal motions onto a suitably constructed subspace that contains specific motions of interest. Examples for proteins are the subspaces of backbone motion, sidechain rotation, bond stretching, *etc.* In order to perform the projection numerically, it is necessary to construct a basis for the subspace of interest. This is a two-step process: first a set of vectors spanning the subspace is selected, and then a corresponding basis is obtained by applying standard algorithms from linear algebra. Note that the initial vectors that span the subspace need not be independent; *any* set of vectors that is known to include all motions of interest can be used. This greatly facilitates the construction of subspaces describing complex motions. Moreover, a small number of motion types are sufficient to cover most situations of practical interest, and we describe these motion types below together with their associated displacement vectors.

We denote the displacement vectors spanning the subspace of interest by $\mathbf{d}_i^{(j)}$, where $i = 1, \ldots, N$ is the atom index and j numbers the vectors. $\mathbf{R}_i$ denotes the position of atom i in the conformation under consideration. A displacement vector without atom index indicates the full $3N$-dimensional vector in configuration space.

- Arbitrary motions of a subset of n atoms labeled $k_1, \ldots, k_n$, for example all backbone atoms or all hydrogen atoms:

$$\begin{aligned} \mathbf{d}_i^{(3j-2)} &= \mathbf{e}_x \delta_{ik_j}, \\ \mathbf{d}_i^{(3j-1)} &= \mathbf{e}_y \delta_{ik_j}, \\ \mathbf{d}_i^{(3j)} &= \mathbf{e}_z \delta_{ik_j}, \quad j = 1, \ldots, n. \end{aligned} \tag{1}$$

  Each of the $3n$ vectors of length $3N$ thus has exactly one component which is one, all other components being zero.
- Rigid-body translation of one group of n atoms labeled $k_1, \ldots, k_n$, for example an amino acid sidechain or a water molecule:

$$\begin{aligned} \mathbf{d}_i^{(1)} &= \mathbf{e}_x \sum_{l=1}^{n} \delta_{ik_l}, \\ \mathbf{d}_i^{(2)} &= \mathbf{e}_y \sum_{l=1}^{n} \delta_{ik_l}, \\ \mathbf{d}_i^{(3)} &= \mathbf{e}_z \sum_{l=1}^{n} \delta_{ik_l}. \end{aligned} \tag{2}$$

Here each vector has several components of value one, one for each atom, indicating that all atoms move together.

- Rigid-body rotation of one group of n atoms labeled $k_1, \ldots, k_n$, for example an amino acid sidechain or a water molecule:

$$\mathbf{d}_i^{(1)} = \sum_{l=1}^{n} \delta_{ik_l} \mathbf{e}_x \times (\mathbf{R}_{k_l} - \mathbf{R}^{(0)}),$$

$$\mathbf{d}_i^{(2)} = \sum_{l=1}^{n} \delta_{ik_l} \mathbf{e}_y \times (\mathbf{R}_{k_l} - \mathbf{R}^{(0)}),$$

$$\mathbf{d}_i^{(3)} = \sum_{l=1}^{n} \delta_{ik_l} \mathbf{e}_z \times (\mathbf{R}_{k_l} - \mathbf{R}^{(0)}) \tag{3}$$

$\mathbf{R}^{(0)}$ is the reference point for the rotation; if rigid body translation is also included in the subspace under consideration, the value of $\mathbf{R}^{(0)}$ does not matter and it can be set to zero.

- Motion of atoms j and k along their distance vector, for example bond stretching:

$$\mathbf{d}_i = (\mathbf{R}_j - \mathbf{R}_k)(\delta_{ij} - \delta_{ik}) \tag{4}$$

Other motions, *e.g.*, bond angle bending, could be added, but are rarely necessary. Combined bond stretching and bond angle bending can be described by including all three interatomic distances involved in a bond angle *via* Eq. (4), and the effect of bond angle bending alone can be studied by comparing this combination to a set of vectors describing only the bond stretching. Some other motion subspaces are most conveniently described by the motions that they do *not* contain. A good example is the torsion angle space, which is most easily defined as the orthogonal complement to the space of bond stretching and bond angle bending. This approach avoids the complicated mathematical expressions inevitably associated with non-Cartesian coordinates. If required, the displacement vector associated with a general coordinate $q(\mathbf{R}_1, \ldots, \mathbf{R}_N)$ can always be derived from the general relation

$$\mathbf{d}_i = \frac{\partial}{\partial \mathbf{R}_i} q(\mathbf{R}_1, \ldots, \mathbf{R}_N). \tag{5}$$

However, it should be pointed out that not all displacement vectors can be written as derivatives of a general coordinate, or at least not easily; the rigid-body rotation in Eq. (3) is a well-known example, and a more complex example will be shown in Section 3.3.

Once all vectors describing motions of interest are collected, they are transformed into a basis spanning the same subspace. This involves an orthonormalization and the elimination of redundant vectors. This step is necessary in order to perform the projection of motion vectors onto the subspace of interest efficiently. As is well-known from linear algebra, a basis constructed from a set of vectors spans the same subspace and as these vectors. There is thus no modification of the physical content of the set of motion vectors.

Several numerical algorithms are available for orthonormalization, of which the best known is the Gram-Schmidt method, which is however not numerically stable. A better method is singular value decomposition (SVD) [14] which is implemented in many standard numerical libraries; we use the LAPACK library [15]. If the system to be analyzed contains N atoms, the input to SVD is the $3N \times M$ matrix

$$\mathbf{D} = (\mathbf{d}^{(1)}, \ldots, \mathbf{d}^{(M)}) \tag{6}$$

whose columns are the M vectors $\mathbf{d}^{(i)}$ spanning the subspace of interest. In the following we call this subspace $\mathcal{T}$. SVD decomposes the matrix $\mathbf{D}$ as

$$\mathbf{D} = \mathbf{U} \cdot \mathbf{\Sigma} \cdot \mathbf{V}^T, \tag{7}$$

where $\mathbf{U}$ is a $3N \times M$ matrix whose columns are orthonormal and $\mathbf{V}$ is an orthogonal $M \times M$ matrix. The matrix $\mathbf{\Sigma}$ is diagonal, $\mathbf{\Sigma} = \mathrm{diag}(\sigma_1, \ldots, \sigma_M)$, and contains the singular values of $\mathbf{D}$, *i.e.*, the eigenvalues of $\mathbf{D}^T \cdot \mathbf{D}$. The number of non-zero singular values, f, is the rank of $\mathbf{D}$ and thus the dimension of $\mathcal{T}$. It can be smaller than M if the input vectors are not linearly independent. The first f columns of $\mathbf{U} = (\mathbf{u}^{(1)}, \ldots, \mathbf{u}^{(M)})$ correspond to non-zero singular values and form an orthonormal basis of $\mathcal{T}$. Most SVD implementations can also provide a basis for the f'-dimensional orthogonal complement $\mathcal{T}'$ at little extra computational cost, where $f + f' = 3N$. As has been mentioned above, this is a useful feature for subspaces which are most easily described *via* their orthogonal complement.

Consider now a given $3N$-dimensional infinitesimal displacement vector $\mathbf{v}$ that we wish to analyze and which we assume to be normalized ($|\mathbf{v}|^2 = 1$) for convenience. Introducing orthonormal basis vectors $\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(f')}$ spanning $\mathcal{T}'$, we can define the two projections

$$\mathbf{v}_{\mathcal{T}} = \sum_{k=1}^{f} (\mathbf{v} \cdot \mathbf{u}^{(k)}) \mathbf{u}^{(k)},$$

$$\mathbf{v}_{\mathcal{T}'} = \sum_{l=1}^{f'} (\mathbf{v} \cdot \mathbf{w}^{(l)}) \mathbf{w}^{(l)}. \tag{8}$$

Since $|\mathbf{v}_T|^2 + |\mathbf{v}_{T'}|^2 = 1$, they represent the contributions of the motion to $\mathcal{T}$ and $\mathcal{T}'$, respectively.

3. APPLICATION TO PROTEIN NORMAL MODES

We have applied the techniques described in the last section in order to study which kinds of motion occur in low-frequency normal modes of proteins. The three example proteins we have used are crambin (PDB entry 1CBN), lysozyme (PDB entry 135L), and myoglobin (PDB entry 1MBD). Larger proteins could not be studied due to the memory requirements of the normal mode calculations and the subsequent analysis. All calculations were performed using the Molecular Modeling Toolkit [16]; the new analysis techniques were implemented in a combination of Python and C code. All proteins were treated in vacuum using the Amber 94 force field [17] without any cutoff for the non-bonded interactions. The experimental structures were minimized up to a remaining energy gradient of 10^{-4} kJ/mol/nm using MMTK's conjugate gradient minimizer.

3.1. Vibrational Density-of-states

The quantity of interest for our analysis is the vibrational spectrum which is usually called the density-of-states (DOS). The DOS is a histogram of the frequencies in the system which is normalized such that the integral over all frequencies is one. We calculated the DOS using a bin width of 0.6 THz. Each value was then replaced by the average of itself and its two neighbors in order to obtain a smoother curve. Although these steps are commonly used for purely technical reasons (*i.e.*, obtaining a useful graphical representation), they also have a physical motivation. Strictly speaking a normal mode spectrum is discrete, and the density of states is a sum of delta functions,

$$g(\omega) = \frac{1}{3N} \sum_{\lambda=1}^{3N} \delta(\omega - \omega_\lambda). \tag{9}$$

Here ω_λ are the eigenfrequencies of the system, where we also include the zero frequencies corresponding to global rotations and translations. A more general definition of $g(\omega)$ which does not depend on a dynamical model can be given in terms of the mass-weighted velocity autocorrelation functions $\langle \mathbf{v}_i(0) \cdot \mathbf{v}_i(t) \rangle$ $(i = 1, \ldots, N)$ of the particles [19]:

$$g(\omega) = \frac{1}{6\pi N k_B T} \sum_{j=1}^{N} \int_{-\infty}^{\infty} dt \exp[-i\omega t] m_j \langle \mathbf{v}_j(0) \cdot \mathbf{v}_j(t) \rangle. \tag{10}$$

As usual, k_B is the Boltzmann constant and T the temperature. The discrete nature of the normal mode spectrum is due to the complete absence of friction, which would in a more realistic model be caused by diffusive motions in the solvent and in the protein itself. Adopting a model where each normal mode can be assigned a friction constant, the effect on $g(\omega)$ would be a broadening of each delta-shaped peak into a Lorentzian with a width proportional to the mode's friction constant. The bin width we have chosen has the right order of magnitude to be considered as a crude way of including friction effects, as can be seen by comparison with a Langevin mode study [18]. We note at this point that expression (10) is used to compute $g(\omega)$ from Molecular Dynamics simulations.

The full frequency spectrum of the three proteins is shown in Figure 1. It is evident that the spectra are very similar; the small differences can be attributed to a different ratio of the twenty amino acids and the different secondary structure. The very-low-frequency domain motions that are

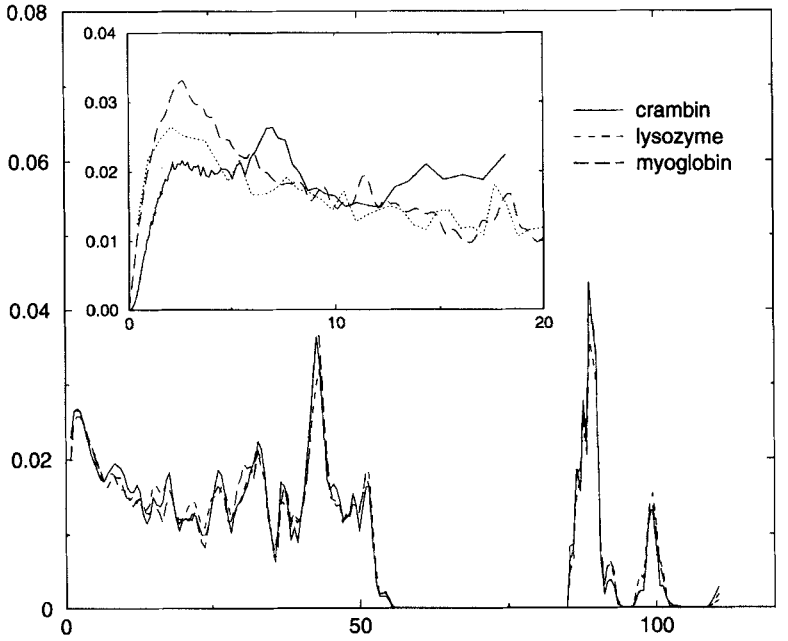

FIGURE 1 The frequency spectrum of crambin, lysozyme, and myoglobin in comparison. They are very similar, suggesting that most of the motions are not specific for a particular protein. The inset shows the low-frequency part of the spectrum (dotted line) in comparison with neutron scattering data from Ref. [21] (solid line) and molecular dynamics results from Ref. [5], (dashed line) both for myoglobin as well. This comparison shows that the spectra are similar enough to permit the application of information gained from normal mode analysis to other techniques.

specific to each protein are so few in number that they are not visible in the histogram. The very-high-frequency part of the spectrum from 80 to 120 THz is known to describe bond vibrations involving hydrogen atoms [20] and is separated from the lower frequency band by a wide gap due to the small mass of the hydrogen atoms; it will not be shown in the following in order to focus on the more interesting lower frequency region. The inset in Figure 1 shows the low-frequency part of the spectrum for myoglobin in comparison with neutron scattering data ("dry myoglobin" from Fig. 5 in [21]) and molecular dynamics results for myoglobin in vacuum (Fig. 6 in [5]) using the united-atom version of the CHARMM force field [22] and a distance-dependent dielectric function $\varepsilon(r) = \varepsilon_0 \cdot (r/r_0)$, with $r_0 = 1$ Å. At this point it should be mentioned that the experimentally measured DOS for myoglobin is practically identical with the DOS of the hydrogen atoms. This is due to the fact that incoherent neutron scattering from hydrogen dominates by far all other scattering processes [23]. Considering the inclusion of anharmonic effects in the molecular dynamics simulation and the use of different force fields, the two theoretical curves are remarkably similar, indicating that the normal mode approximation is not unrealistic for our study. The neutron scattering data shows much more important differences, which are due to both deficiencies in the theoretical model and various difficulties in obtaining the experimental spectrum. Nevertheless, the comparison shows that the frequency scales are essentially the same, allowing an application of our interpretations to spectral data of different origin.

For showing the various decompositions of the normal modes, we calculate the contribution of a specific subspace to each normal mode and then a weighted histogram of the vibrational frequencies into which each frequency enters with a weight that is equal to the contribution of the subspace to the corresponding mode. The results are again very similar for the three proteins we have studied, and therefore we will demonstrate each contribution for only one protein.

3.2. Low-frequency Rigid-body Motions

Our first goal is the interpretation of the low-frequency region of the spectrum, and especially the first peak at a frequency of ≈ 2 THz or 60 cm^{-1}. A previous study of molecular dynamics trajectories [5, 6] found that both diffusional and vibrational motion in this frequency range can be described by a rigid sidechain model, and that the diffusive motion is dominated by rigid sidechain diffusion. It is therefore interesting to see if a similar

assignment can be made for vibrational motions in the harmonic approximation. Using the methods described in Section 2, we calculated the contributions of two subspaces to each of the normal modes: the subspace of backbone motions, and the combined subspace of backbone motions and rigid-body motions of sidechains. The result is shown in Figure 2a for myoglobin. The area below the backbone line represents the backbone contribution, the area between the backbone line and the backbone/rigid sidechain line represents the sidechain rigid-body contribution, and the remaining area represents the internal motions of the sidechains. Three quarters of the first peak are indeed contributed by rigid-body motions of the sidechains, whereas one quarter comes from backbone motions and almost nothing from the internal sidechain motions. The lowest frequency band with sidechain deformation lies between 5 and 17 THz (150 to 510 cm^{-1}) and describes the collective deformations of the sidechains. In the frequency range above 25 THz (750 cm^{-1}), the contributions from backbone

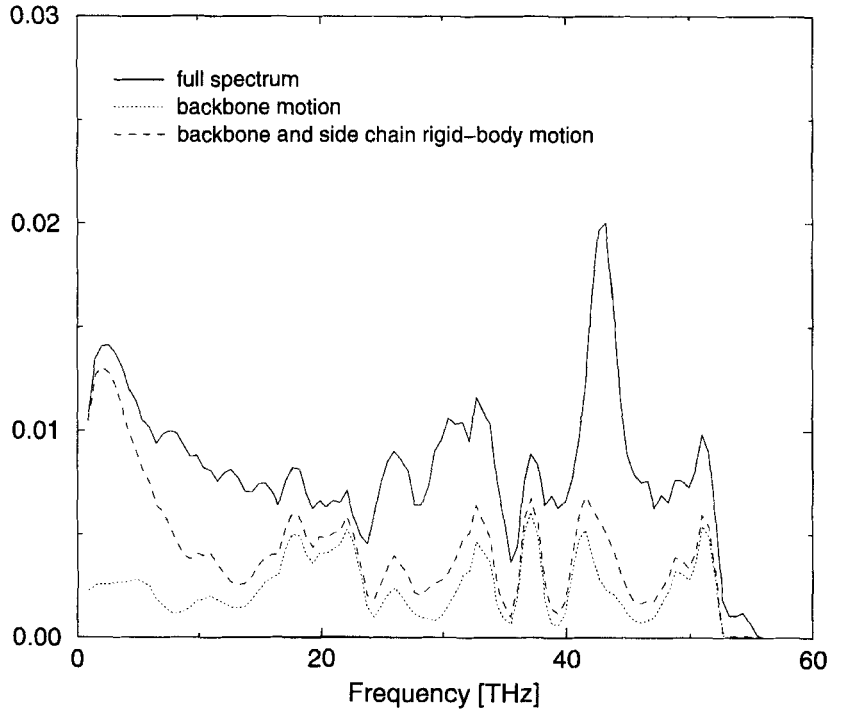

FIGURE 2a The frequency spectrum of myoglobin, and the contributions from backbone motion and backbone plus rigid-body sidechain motion. The dominant contribution to the first peak comes from the rigid-body motion of the sidechains.

and internal sidechain motion have essentially similar shape and a roughly constant ratio equal to the ratio of sidechain to backbone atoms in the protein, which is 4:3. These regions thus represent small-scale vibrations of chemical units that occur equally in backbone and sidechains.

The observation that the first peak in the frequency spectrum is mainly caused by rigid-body motions of the sidechains does not by itself characterize this frequency range very well. Rigid-body motions of sidechains include both large-scale collective motions of multiple residues, in which the sidechains participate just like the backbone atoms, and sidechain motion relative to their immediate surrounding. The latter are essentially rotations, because sidechains are attached to the backbone by a stiff chemical bond that causes relative translations to have a rather high frequency. Figure 2b

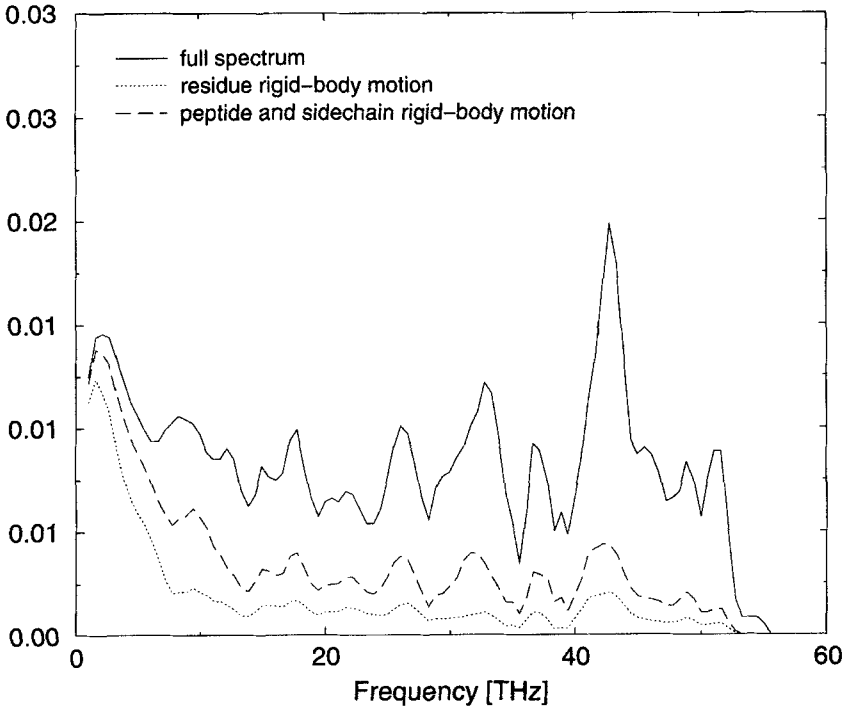

FIGURE 2b The frequency spectrum of crambin, and the contributions from rigid-body motion of the individual residues, as well as the contribution from a separate rigid-body motion of the sidechain and peptide parts of each residue. The area between these two curves thus represents the contribution from sidechain rotation relative to the backbone, whereas the area below the lowest curve represents collective residue motions, *i.e.*, deformations at the level of the secondary and tertiary structure.

shows another decomposition of the frequency spectrum that permits a more detailed analysis of sidechain motions. The lowest curve shows the contribution of rigid-body motion of the entire residues to the spectrum. This contribution contains all the large-scale motions above the residue level, *i.e.*, anything from secondary-structure dynamics to domain motions. As expected, it is limited to very low frequencies (up to ≈ 7 THz ($210\,\mathrm{cm}^{-1}$)), and constitutes the most important part of the first peak. The middle curve adds the contribution of the motion of the sidechains relative to the peptide part of their residue. This term has almost the same frequency distribution as the first one, showing that sidechain rotations occupy essentially the same frequency range as full-residue motions. This indicates a strong coupling between rotations and translations in the vibrational rigid sidechain dynamics. It is also visible that rotational rigid sidechain motion contributes little to the total vibrational spectrum.

3.3. Spatial Frequency Decomposition

It is a well-known general feature of physical systems that fast motions are localized, involving only a few atoms that are close to each other, whereas slow motions describe large-scale deformations. This behavior is caused by the distance dependence of the relevant interactions: interactions between atoms at short distances, *e.g.*, bonded atoms, are much stronger than the smooth long-range interactions between distant atoms. Mathematically it is expressed by a generally monotonic relation between spatial and temporal frequencies, known as a dispersion relation. A standard dispersion relation is not meaningful for proteins, because they are specific finite-size inhomogeneous objects. However, it is still of interest to study the relation between spatial and temporal frequencies for elastic vibrations of proteins.

Since normal mode analysis provides the motions of a protein sorted by temporal frequency, all that remains to be done is to analyze each mode according to its spatial frequency content. This can be achieved in much the same way as the geometrical analyses described in the last section: one constructs a basis for the subspace of all motions up to a given spatial frequency, and calculates the projection of each normal mode vector onto this subspace. This subspace provides a rather complex example of motions that can only be treated infinitesimally; there are no generalized coordinates of which the displacement vectors shown below are derivatives. Moreover, this example illustrates an approach that could be of more general use, namely the definition of displacement vectors from vector fields in the $3N$-dimensional Cartesian space.

The construction of a basis for this subspace has been described in detail in Ref. [24], where it has been used to obtain low-frequency normal modes efficiently, and will only be summarized here. It is based on the idea that a set of atomic displacement vectors can be regarded as the values of a vector field $\mathbf{D}(\mathbf{r})$, defined everywhere in space, at the positions of the atoms, *i.e.*, $\mathbf{d}_i = \mathbf{D}(\mathbf{R}_i)$, where $\mathbf{R}_i$ is the position of atom i and $\mathbf{d}_i$ is its displacement vector. A basis for a subspace of infinitesimal motions can thus be obtained from a suitable collection of vector fields. For a wavelength-dependent subspace, these vector fields must take the form of sine or cosine waves, with wave numbers up to the specified cutoff. Such waves are conveniently constructed inside a rectangular box around the protein which is replicated periodically in space. This box has no physical reality; its purpose is to define a lower limit to the wave numbers that are included in the subspace. Artifacts due to the periodicity can be avoided if the box is sufficiently large; it must exceed the minimal bounding box around the protein by at least half the minimal wavelength.

A precise specification of the basis for the vector field $\mathbf{D}(\mathbf{r})$ is given by the set of vector fields

$$\mathbf{B}_\alpha^{ijk}(\mathbf{r}) = w(x, k_i^{(x)}) w(y, k_j^{(y)}) w(z, k_k^{(z)}) \mathbf{e}_\alpha, \tag{11}$$

where $\mathbf{e}_\alpha, \alpha = x, y, z$ is a unit vector along one of the three Cartesian axes and

$$w(x,k) = \begin{cases} \sin(kx) & \text{for } k < 0 \\ \cos(kx) & \text{for } k \geq 0 \end{cases} \tag{12}$$

The wavenumbers are given by

$$k_i^{(\alpha)} = \frac{2\pi}{L_\alpha} n_i, \tag{13}$$

where n_i is an integer and L_α is the length of the enclosing box along coordinate axis α. The total set of wavenumbers to be used is defined by the condition

$$\sqrt{k_i^{(x)^2} + k_j^{(y)^2} + k_k^{(z)^2}} < \frac{2\pi}{\lambda_{\min}} \tag{14}$$

To construct a set of basis vectors from the vector fields $\mathbf{B}_\alpha^{ijk}(\mathbf{r})$, the first step is the conversion of each vector field into a set of atomic displacement vectors, *i.e.*, $\mathbf{B}_\alpha^{ijk}(\mathbf{R}_l)$. This set is then orthonormalized as described in Section 2.

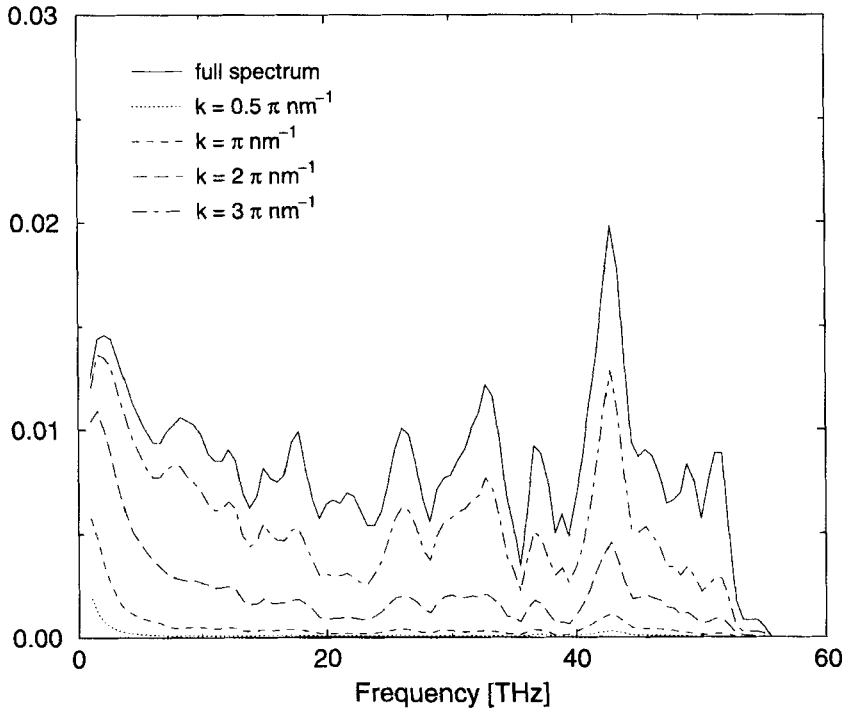

FIGURE 3a Spatial frequency contributions to the normal modes of crambin. Each curve represents the contributions of all wave numbers up to the limiting value indicated in the plot. It is clearly visible that collective motions (low wave numbers) correspond to slow motions, but that no clear relation between spatial and temporal frequencies exists for higher frequencies.

The normal modes of crambin have been analyzed according to their spatial frequency content as described in Section 3.3. Figure 3a shows the contributions for four different spatial frequency cutoffs. It is clear that up to $k \approx \pi\,\mathrm{nm}^{-1}$, corresponding to a length scale of 2 nm, only very slow motions occur; these motions include everything from helix deformations to domain motions. The curve at $k = 2\pi\,\mathrm{nm}^{-1}$ already contains significant contributions at all frequencies. This shows that at the corresponding length scale (≈ 1 nm) the detailed chemical structure of the protein becomes visible, and a simple relation between spatial and temporal frequencies like in continuous media can no longer be expected. It should be noted that at $k \approx 4\pi\,\mathrm{nm}^{-1}$ (≈ 0.5 nm) the full spectrum is reached, because smaller length-scale motions do not exist.

The spatial frequency decomposition was further evaluated on a much finer wavenumber grid (in steps of $0.125\pi\,\mathrm{nm}^{-1}$) in order to permit the

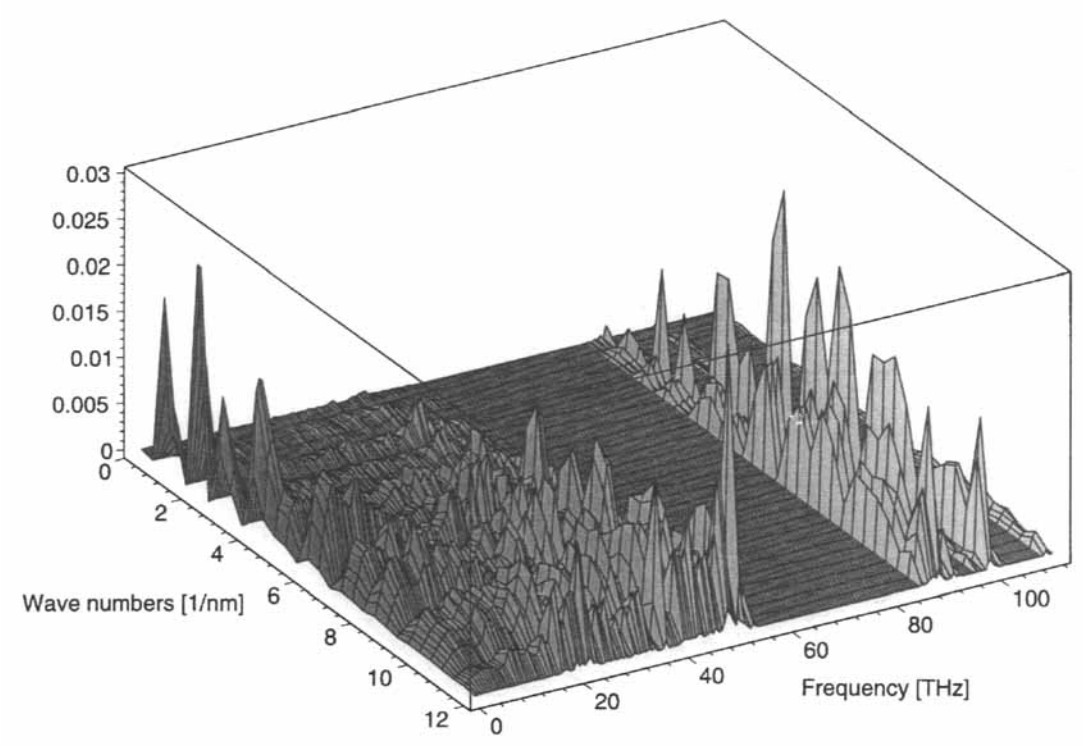

FIGURE 3b Spatial-temporal frequency spectrum of crambin (in arbitrary units). The dominance of slow motions at low wave numbers is evident; the discrete peak structure in that region is a consequence of the finite system size, as discussed in the text. The extension of the well-separated fast-motion band to rather low wave numbers shows that these motions (hydrogen bond stretching) do not fit precisely into a collective-motion description.

calculation of a two-dimensional spectrum by numerical differentiation. This spectrum is shown in Figure 3b. Again it is evident that up to $k \approx 3\,\mathrm{nm}^{-1}$ only very low frequency motions are possible. The discrete structure of the spectrum in this region is a consequence of the finite size of the system; in between the peaks that are situated periodically along the k-axis there are no additional wavenumber vectors that could contribute. With increasing k, the low-frequency motions contribute less and the higher frequencies become more important. Of particular interest is the well-separated band describing the bond stretching motions of the hydrogen atoms. It starts to show important contributions at unexpectedly low k values ($\approx 2\,\mathrm{nm}^{-1}$), much lower than the onset of some other motions which have substantially lower frequencies. This shows that the hydrogen motions are not well described by a uniform collective motion picture. This behavior is reminiscent of the normal mode structure of solids, the low-frequency band corresponding to the acoustic branch (for $k \to 0$ we find $\omega \to 0$ as well), and the high-frequency band corresponding to an optical branch (almost reaching $k = 0$ but remaining at a finite frequency).

It is also interesting to examine the frequency region below 2 THz. Although the principal contributions come from the peaks at low wavenumbers, which describe collective motions, there is a clearly non-zero contribution from higher wavenumbers as well. They indicate that at a sufficiently small length scale, a protein is not a homogeneous material. This is consistent with the observation that sidechain motions in this frequency range can be described by rigid-body motions; each sidechain has somewhat different rigid-body motion parameters than its neighbors, and these relative motions between the almost rigid sidechains create the high-wavenumber contributions at low frequencies. The spatial frequency decomposition alone thus already contains information about the length scale of inhomogeneities in the protein.

4. CONCLUSION

We have presented a new technique for analyzing complex motions in macro-molecules by projection on particular subspaces. This technique can be applied to the analysis of any kind of motion which can be considered to some degree as 'infinitesimal', *e.g.*, vibrations or fluctuations around an equilibrium configuration. The basic input is any set of displacement vectors describing the physical situation one wishes to study. In this way a large set of particular motion types can be considered, ranging from motions of

subsets of the system to collective motions defined *via* vector fields. The subspace basis for projecting the motions is then constructed by singular value decomposition.

Our application of this technique to protein normal modes leads to the following general picture of protein dynamics: All motions above $\approx 20\,\text{THz}$ ($600\,\text{cm}^{-1}$) are small-scale vibrations that are due to the chemical structure of the individual amino acid residues that make up every protein. Between 10 and 20 THz (300 to $600\,\text{cm}^{-1}$), we see motions that involve both the covalent bond structure and non-bonded interactions; this range includes the collective deformations of the longer sidechains. The first peak in the frequency spectrum, around 2 THz ($60\,\text{cm}^{-1}$), describes secondary-structure motions on length scales of 1 nm and more, *e.g.*, helix deformations. For frequencies below 2 THz, the protein sidechains can be considered as rigid objects, confirming earlier findings. Our analysis has shown that rotational rigid sidechain motion does not give important contributions to the vibrational density of states and is strongly coupled to the motion of the backbone. Comparing to earlier work one can also conclude that diffusive motion in proteins is mostly due to global rotational motions of the sidechains. The large-scale collective motions, *e.g.*, domain motions, occupy an invisibly small part at the extreme lower end of the spectrum; it has been shown previously that they depend only on the three-dimensional structure of a protein plus the existence of some unspecific mid-range interactions that become weaker with increasing distance.

Acknowledgement

We thank Dr. W. Doster for providing the neutron scattering data from Ref. [21] used in Figure 1.

References

[1] Smith, J. C. (1991). "Protein dynamics: comparison of simulations with inelastic neutron scattering experiments", *Q. Rev. Biophys.*, **24**, 227–291.
[2] Smith, J. C. and Kneller, G. R. (1993). "Combination of neutron scattering and molecular dynamics to determine internal motions in biomolecules", *Mol. Sim.*, **10**(2–6), 363–375.
[3] Kneller, G. R. and Geiger, A. (1989). "A method to calculate the *g*-coefficients of the molecular pair correlation function from molecular dynamics simulations", *Mol. Sim.*, **3**, 283–300.
[4] Kneller, G. R. (1991). "Superposition of molecular structures using quaternions", *Mol. Sim.*, **7**, 113–119.
[5] Furois-Corbin, S., Smith, J. C. and Kneller, G. R. (1993). "Picosecond timescale rigid-helix and side-chain motions in deoxymyoglobin", *Proteins*, **16**, 141–154.
[6] Kneller, G. R. and Smith, J. C. (1994). "Liquid-like sidechain dynamics im myoglobin", *J. Mol. Biol.*, **242**, 181–195.

[7] Go, N., Noguti, T. and Nishikawa, T. (1983). "Dynamics of a small protein in terms of low-frequency vibrational modes", *Proc. Nat. Acad. Sci. U.S.A.*, **80**, 3696–3700.
[8] Lagant, P., Vergoten, G., Fleury, G. and Loucheux-Lefebvre, M.-H. (1984). "Raman spectroscopy and normal vibrations of peptides", *Eur. J. Biochem.*, **139**, 137–148.
[9] Schulz, G. E. (1991). "Domain motions in proteins", *Curr. Opin. Struct. Biol.*, **1**, 883–888.
[10] Hinsen, K., Thomas, A. and Field, M. J. (1999). "Analysis of domain motions in large proteins", *Proteins*, **34**, 369–382.
[11] Smith, J. C., Cusack, S., Pezzeca, U., Brooks, B. and Karplus, M. (1986). "Inelastic neutron scattering analysis of low frequency motion in proteins: A normal mode study of the bovine pancreatic trypsin inhibitor", *J. Chem. Phys.*, **85**, 3636–3654.
[12] Smith, J. C., Cusack, S., Tidor, B. and Karplus, M. (1990). "Inelastic neutron scattering analysis of low-frequency motions in proteins: Harmonic and damped harmonic models of bovine pancreatic trypsin inhibitor", *J. Chem. Phys.*, **93**, 2974–2991.
[13] Kitao, A. and Go, N. (1999). "Investigating protein dynamics in collective coordinate space", *Curr. Opin. Struct. Biol.*, **9**, 164–169.
[14] Press, W. H., Teukolsky, S. A., Vetterling, W. T. and Flannery, B. P., Numerical Recipes in C, 2nd edition, Cambridge University Press, Cambridge, 1992.
[15] Anderson, E., Bai, Z., Bischof, C., Demmel, J. W., Dongarra, J. J., du Croz, J., Greenbaum, A., Hammarling, S., McKenney, A. and Sorensen, D., "LAPACK: A portable linear algebra library for high-performance computers", Computer Science Dept. Technical Report CS-90-105, University of Tennessee, Knoxville, 1990.
[16] Hinsen, K., "The Molecular Modeling Toolkit: A case study of a large scientific application in Python", *Proceedings of the 6th International Python Conference*, http://www.python.org/workshops/1997–10/proceedings/hinsen.html
[17] Cornell, W. D., Cieplak, P., Bayly, C. I., Gould, I. R., Merz, K. M. Jr., Ferguson, D. M., Spellmeyer, D. C., Fox, T., Caldwell, J. W. and Kollman, P. A. (1995). "A second generation force field for the simulation of proteins and nucleic acids", *J. Am. Chem. Soc.*, **117**, 5179–5197.
[18] Ansari, A. (1999). "Langevin modes analysis of myoglobin", *J. Chem. Phys.*, **110**, 1774–1780.
[19] Rahman, A., Singwi, K. S. and Sjölander, A. (1962). "Theory of Slow Neutron Scattering. I", *Phys. Rev.*, **126**, 986–996.
[20] Brooks, C. L., Karplus, M. and Pettit, B. M. (1988). "Proteins, a theoretical perspective of dynamics, structure, and thermodynamics", *Adv. Chem. Phys.*, **71**.
[21] Settles, M. and Doster, W., "Iterative calculation of the vibrational density of states from incoherent neutron scattering data with the account of double scattering", In: *Biological Macromolecular Dynamics*, Eds., Cusack, S., Büttner, H., Ferrand, M., Langan, P. and Timmins, P., Adenine Press, New York, 1997.
[22] Brooks, B. R., Bruccoleri, R. E., Olafson, B. D., States, D. J., Swaminathan, S. and Karplus, M. (1983). "CHARMM: A Program for Macromolecular Energy, Minimization, and Dynamics Calculations", *J. Comp. Chem.*, **4**, 187–217.
[23] Lovesey, S., "Theory of neutron scattering from condensed matter", Vol.1, Clarendon Press, Oxford, 1984.
[24] Hinsen, K. (1998). "Analysis of domain motions by approximate normal mode calculations", *Proteins*, **33**, 417–429.