

Analysis of Domain Motions by Approximate Normal Mode Calculations

Konrad Hinsén*

Laboratoire de Dynamique Moléculaire, Institut de Biologie Structurale—Jean-Pierre Ebel, Grenoble Cedex 1, France

ABSTRACT The identification of dynamical domains in proteins and the description of the low-frequency domain motions are one of the important applications of numerical simulation techniques. The application of these techniques to large proteins requires a substantial computational effort and therefore cannot be performed routinely, if at all. This article shows how physically motivated approximations permit the calculation of low-frequency normal modes in a few minutes on standard desktop computers. The technique is based on the observation that the low-frequency modes, which describe domain motions, are independent of force field details and can be obtained with simplified mechanical models. These models also provide a useful measure for rigidity in proteins, allowing the identification of quasi-rigid domains. The methods are validated by application to three well-studied proteins, crambin, lysozyme, and ATCase. In addition to being useful techniques for studying domain motions, the success of the approximations provides new insight into the relevance of normal mode calculations and the nature of the potential energy surface of proteins. *Proteins* 33:417–429, 1998. © 1998 Wiley-Liss, Inc.

Key words: protein energy surface; crambin; lysozyme; ATCase

INTRODUCTION

In the analysis of protein dynamics, an important goal is the description of slow large-amplitude motions in large proteins. These motions typically describe rearrangements of domains which are essential for the function of the protein. Only such global motions can change the exposed surface of the protein significantly and hence influence interactions with its environment. Higher frequency motions are more localized in the interior or on the surface of the protein. However, this does not mean that they are irrelevant; they can play an important role in signal transmission mechanisms and other internal processes. Indeed, the frequently observed strong influence of single-residue mutations, which are expected to cause only local changes in conformation and dynamics, on protein function indicates a

higher significance of medium frequency motions than is commonly supposed.

One of the standard techniques for studying protein dynamics, and in particular low-frequency domain motions, is normal mode analysis.¹ In contrast to phase space sampling techniques, such as molecular dynamics, normal mode analysis provides a very detailed description of the dynamics around a local energy minimum. The technique has important limitations (harmonic approximation, neglect of solvent damping, no information about energy barriers and crossing events); nevertheless, it has provided much useful insight into protein dynamics. Its most important contribution is the identification and characterization of the low-frequency domain motions. In contrast, the corresponding vibrational frequencies obtained by normal mode analysis are of little physical relevance, because the real frequencies are strongly influenced by anharmonic effects^{2,3} and solvent damping.^{4–6} In fact, low-frequency motions in a realistic environment are overdamped and hence not vibrational at all.

Various studies have shown that the assumption of harmonic motion, which is implicit in normal mode analysis, is justified for medium- and high-frequency modes but not for the slow modes that correspond to domain motion.^{2,3,7} Furthermore, molecular dynamics simulations⁸ and experiments⁹ have shown the existence of conformational substates, corresponding to multiple minima of the potential energy that are accessible by thermal fluctuations. The practical relevance of normal mode analysis therefore seems questionable, because it explores only one specific (and arbitrarily chosen) minimum. Moreover, collective motions at physiological temperatures are not determined by the potential energy surface, but by the potential of mean force expressed as a function of a smaller set of “slow” variables, which is a much smoother function. A comparison of normal modes in several closely spaced local minima of Bovine Pancreatic Trypsin Inhibitor (BPTI)¹⁰ has shown that there is an observable variation of the

Grant sponsor: Human Frontier Science Project Organization.

*Correspondence to: Konrad Hinsén, Centre de Biophysique Moléculaire (CNRS), Rue Charles Sadron, 45071 Orléans Cedex 2, France. E-mail: hinsen@cns-orleans.fr

Received 17 March 1998; Accepted 1 July 1998

low-frequency modes, but this variation occurs within a well-defined subspace. Comparisons of low-frequency normal modes and the directions of large-amplitude fluctuations in molecular dynamics simulations indicate clear similarities.^{3,7} All these observations suggest that dynamical domains and their motions are well defined and can be analyzed using a variety of techniques. On the other hand, the assignment of time scales and amplitudes to these motions requires a detailed model that incorporates anharmonic and solvent effects.

A major practical problem with normal mode analysis is its use of memory ($O(N^2)$, where N is the number of atoms in the protein) and CPU time ($O(N^3)$). Its direct application is therefore limited to small molecules up to approximately 2,000 atoms. Larger systems can be handled by reducing the number of degrees of freedom. A commonly used approximation is the elimination of all bonds and bond angles, leaving only the dihedral angles (or even a subset of all dihedral angles) free to change. However, more severe mechanical constraints such as rigid domains have been proposed,⁷ as well as more complicated coordinate sets that are not simply a subset of the standard internal coordinates.¹¹ The disadvantage of constraint methods is that in addition to eliminating "uninteresting" modes, they modify the remaining ones. Another approach to treating larger systems is the calculation of only the lowest modes. This is achieved by partitioning methods that divide either the physical system according to geometrical criteria¹² or the second derivative matrix by imposing a block structure¹³ into pieces that can be diagonalized exactly. The low-frequency contributions from all parts are then combined to form a reduced basis for treating the complete system. Such methods can deal with very large systems¹⁴ but still require computational resources that are not easily available.

In this article, several physical approximations are presented that permit the calculation of low-frequency modes with a much reduced computational effort. For example, a good approximation to the low-frequency modes of ATCase, whose exact calculation required 690 hours on a Cray C98 supercomputer (Cray Research, Eagan, MN),¹⁴ could be obtained in just 9 minutes on a personal computer. Although approximate, this method is sufficiently accurate to allow the identification of rigid domains and flexible regions in a protein as well as the determination of the principal large-scale motions. However, no effort is made to obtain physically unreliable data accurately, specifically the vibrational frequencies and the highly artificial thermodynamical amplitudes that are commonly derived from them. The success of the approximations also provides further insight into the nature of protein energy surfaces.

The fundamental principle on which all methods presented in this article are based is the fact that low-frequency modes represent global movements of

large domains, whereas high-frequency modes correspond to localized motions involving few atoms. This is a very general observation, which is due to two mechanisms. First, global domain motions have no (or very little) energy contribution from internal degrees of freedom of the domains because there is no deformation. Second, the long-range interactions between domains are weaker than the short-range interactions between neighboring atoms. This principle is used first to find an appropriate small subspace for calculating low-frequency modes, and in a second step to obtain a much simplified force field for approximate normal mode calculations. This simplified force field also provides a straightforward method for identifying dynamical domains by calculating the energy density associated with local deformations due to the normal modes.

METHODS

Normal Modes in a Subspace

To establish the notation for the following sections, the standard normal mode analysis procedure is briefly reviewed here. For details and derivations, the reader should consult textbooks on classical mechanics (e.g., Goldstein¹⁵) and linear algebra.

Normal mode analysis begins with the calculation of the second derivative matrix \mathbf{H} of the potential energy at a local minimum. This matrix is of size $3N \times 3N$, where N is the number of atoms in the molecule. The mass-weighted second derivative matrix is defined by $\mathbf{H}^* = \mathbf{M}^{-1/2} \cdot \mathbf{H} \cdot \mathbf{M}^{-1/2}$, where \mathbf{M} is a diagonal $3N \times 3N$ matrix containing the atomic masses. The normal modes are the eigenvectors of \mathbf{H}^* (in mass-weighted Cartesian coordinates), and the corresponding eigenvalues are the squares of the vibrational frequencies. The limiting factors in normal mode analysis are the memory requirements for storing the matrix \mathbf{H}^* and the CPU time for the eigenvalue calculation.

Normal modes can be calculated in any set of coordinates, not only in the commonly used Cartesian coordinates. To obtain normal modes in a set of coordinates q_i , $i = 1 \dots 3N$, the transformation matrix \mathbf{C} between the differentials of these coordinates and those of the mass-weighted Cartesian coordinates x_i^* , $i = 1 \dots 3N$ must be calculated. It is defined by $C_{ij} = \partial q_i / \partial x_j^*$. The normal modes are then obtained as the eigenvectors of the matrix $\mathbf{H}_q^* = \mathbf{C} \cdot \mathbf{H}^* \cdot \mathbf{C}^T$. The eigenvectors can be transformed back into mass-weighted Cartesian coordinates by multiplying with \mathbf{C}^T .

Any complete set of coordinates will yield the same normal modes. However, it is possible to leave out some of the coordinates q_i , which is physically equivalent to keeping the corresponding degrees of freedom fixed. After a transformation to internal coordinates, for example, it is possible to eliminate bond and bond angle coordinates, leaving only the dihedrals. This results in a smaller matrix \mathbf{H}_q^* , saving memory and

CPU time, but the modes obtained in this way are smaller in number and not identical to the modes obtained from a full normal mode analysis. The goal is to find the smallest subspace that reproduces the low-frequency modes well enough. The most commonly used subspaces contain some combination of dihedral angles. Some less common subspaces are described in Brooks et al.¹¹

In practice the matrix \mathbf{H}_q^* is calculated directly, without first calculating the full Cartesian matrix \mathbf{H}^* . There are several approaches for obtaining \mathbf{H}_q^* : finite difference derivatives along the subspace basis vectors (i.e., the columns of \mathbf{C}), assembly from small parts by multiplying the individual contributions to \mathbf{H} by the corresponding parts of \mathbf{C} , or for short-range force fields storage of \mathbf{H} as a sparse matrix.

Fourier Basis

A dihedral angle subspace, even if limited to the backbone dihedrals ϕ and ψ , is still too large to be used for big proteins. Eliminating more degrees of freedom is possible in principle, but the choice is not obvious and the effect on the modes difficult to predict. Keeping large domains rigid is not a general solution either, because the domains are not a priori known. In addition, although it is common to speak of "rigid" domain movement in low-frequency modes, these domains are not strictly rigid; there is always some overall deformation and intradomain movement. Rigid domains should be viewed as a useful description of protein dynamics rather than as the basis for a mechanical model.

The first step in the construction of a more appropriate subspace for low-frequency mode calculations is the realization that the basis vectors of the subspace are not coordinates but coordinate differentials, i.e., each basis vector describes a direction, not a position, in $3N$ -dimensional coordinate space. A basis vector can therefore be regarded as a set of atomic displacement vectors. A corresponding set of coordinates does not have to be specified; it may not even exist.

It is often useful to treat a set of atomic displacement vectors as the values of a vector field, which is defined everywhere in space, at the positions of the atoms, i.e.,

$$\mathbf{d}_i = \mathbf{D}(\mathbf{R}_i), \quad (1)$$

where \mathbf{R}_i is the position of atom i and \mathbf{d}_i is its displacement vector. Obviously, there is more than one vector field $\mathbf{D}(\mathbf{r})$ corresponding to a given set of displacement vectors \mathbf{d}_i , although the inverse relation is unique. Because the vector field $\mathbf{D}(\mathbf{r})$ has no direct physical meaning, this is not a problem. If a vector field is to be constructed from a set of displacement vectors, e.g., for analysis or visualization, the most reasonable choice is a field that varies smoothly between the atoms.

A basis for normal mode calculations can thus be obtained from a complete set of functions defined in a region of space that includes the whole protein. An appropriate function set for separating localized from nonlocalized motions is a collection of sine and cosine functions, defined in a rectangular box enclosing the protein. Because this is an infinite function set, a lower limit must be set for the wavelengths λ to be included. The shortest distance over which completely independent motion will be permitted by such a basis is $\lambda/2$. Obviously the smallest reasonable wavelength is thus twice the smallest interatomic distance; this would result in a basis equivalent to the full Cartesian basis. A larger cutoff wavelength leads to a smaller basis which still includes the interesting low-frequency motions. Care must be taken to avoid artifacts resulting from the periodicity of the function set; the box enclosing the protein must be larger than a minimal bounding box by half the cutoff wavelength.

A precise specification of this normal mode subspace basis is given by the vector fields

$$\mathbf{B}_\alpha^{ijk}(\mathbf{r}) = w(x, k_i^{(x)})w(y, k_j^{(y)})w(z, k_k^{(z)})\mathbf{e}_\alpha, \quad (2)$$

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

$$w(x, k) = \begin{cases} \sin(kx) & \text{for } k < 0 \\ \cos(kx) & \text{for } k \geq 0. \end{cases} \quad (3)$$

The wavenumbers are given by

$$K_i^{(\alpha)} = \frac{2\pi}{l_\alpha} n_i, \quad (4)$$

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

$$\sqrt{K_i^{(x)^2} + K_j^{(y)^2} + K_k^{(z)^2}} < \frac{2\pi}{\lambda_{\min}}. \quad (5)$$

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 according to equation 1. Then three basis vectors describing the global rotation of the protein are added; this is advisable because the global rotation of the protein is not well represented by a Fourier basis unless the cutoff wavelength is very small. The complete basis set is then converted to mass-weighted coordinates and orthogonalized, e.g., by singular value decomposition.

The scheme described above allows the construction of a normal mode basis for any system and with

almost arbitrary size. It is therefore easily applicable to other macromolecular systems as well.

Simplified Force Field

Because normal mode analysis requires the evaluation of the potential energy and its first and second derivatives for a single configuration only, the computational cost of this step is normally negligible. However, the force field influences the total computational cost for a normal mode analysis indirectly: a careful energy minimization is required before the second derivative matrix is calculated. Depending on the method used for calculating the second derivative matrix in the chosen basis set, other cost factors may be given by numerical (finite difference) differentiation and/or storage of the nonsparse Cartesian second derivative matrix for long-range force fields.

Although a strong influence of force field details (such as electrostatic cutoff) on the lowest vibrational frequencies¹⁶ has been observed, it can be expected that the distinction between low- and high-frequency modes depends much more on the global versus local character of the deformations than on the precise functional form of the force field. It is therefore reasonable to attempt a normal mode analysis with a much simplified force field.

The functional form used in this work is

$$U(\mathbf{R}_1, \dots, \mathbf{R}_N) = \sum_{\text{all pairs } i,j} U_{ij}(\mathbf{R}_i - \mathbf{R}_j) \quad (6)$$

with an harmonic pair potential

$$U_{ij}(\mathbf{r}) = k(\mathbf{R}_{ij}^{(0)}) (|\mathbf{r}| - |\mathbf{R}_{ij}^{(0)}|)^2, \quad (7)$$

where $\mathbf{R}_{ij}^{(0)}$ is the pair distance vector $\mathbf{R}_i - \mathbf{R}_j$ in the input configuration. In other words, the force field is constructed on the basis of the *assumption* that the input configuration corresponds to a local minimum. An energy minimization is therefore not necessary, but of course the input configuration must be physically reasonable, which can be assumed for structures obtained by X-ray crystallography or nuclear magnetic resonance (NMR). The pair force constant could be given by any function that decreases with distance; for practical reasons the form

$$k(\mathbf{r}) = c \cdot \exp\left(-\frac{|\mathbf{r}|^2}{r_0^2}\right) \quad (8)$$

was chosen; the exponential decay allows the evaluation with a cutoff not significantly larger than r_0 , and the quadratic dependence on \mathbf{r} eliminates the need for a relatively expensive square root calculation. The distance r_0 was set to 0.3 nm; this value gives the best agreement for the low-frequency modes of lysozyme with the Amber force field (this comparison will be shown below). It does not have to be determined

accurately, because the normal modes do not depend strongly on it. The value of c is arbitrary, because it causes only a uniform scaling of all vibrational frequencies.

Initially, the force field described above seems much too simple to describe proteins. It does not take into account elemental features such as the bond structure or the chemical elements of the atoms. However, this force field will be applied only to protein configurations that are known to be physically reasonable, e.g., crystallographic structures or configurations obtained by modeling techniques. In real protein configurations, the various energy terms from a standard empirical force field correspond to different interatomic distances (atoms linked by a bond are closer than atoms that interact via a dihedral angle term only), and the interaction strength decreases with increasing distance. This important feature is captured by the distance-dependent force constant given in equation 8.

A simple harmonic force field similar to the one presented here has been used by Tirion¹⁷ to reproduce the density of slow vibrational modes in proteins. It uses a different distance dependence of the pair force constant, namely a constant up to a certain cutoff distance (which depends on the van der Waals radii of the two atoms involved) and zero beyond that distance. That choice reproduced the density of modes well, but no comparison of the normal mode displacements was shown. The pair force constant given in equation 8 seems more realistic and is easier to apply, because it depends on no specific atom parameters.

Simplified Protein Model

Despite the reduction of memory and CPU time requirements by use of a Fourier basis, the system size is still severely limited by available memory. The largest matrix that must be stored is usually the set of basis vectors, whose size is $M \times 3N$, where N is the number of atoms and M is the number of modes to be calculated. Reducing M by increasing λ_{\min} leads to less accurate modes. When λ_{\min} is already much larger than typical interatomic distances, it is more reasonable to reduce the amount of detail in the protein model rather than increase the cutoff wavelength.

An obvious simplified model for proteins consists of one point mass per residue, located either at the center of mass of the residue or at the C_α position. Such a model is sufficient to study backbone motion, which in turn is sufficient to characterize the low-frequency modes of large proteins. The main difficulty with simplified protein models is the need to construct an appropriate force field; for normal mode analysis, however, the simplified force field described in the previous section can easily be adjusted to such models. For the residue point mass model, the distance r_0 in equation was increased to 0.7 nm,

the value that gives the best agreement for the normal modes of ATCase (discussed below).

Such a model can be interpreted as representing the potential of mean force as a function of the residue positions. Obviously the real potential of mean force is a much more complicated function, and its form is essentially unknown. However, for the purpose of characterizing domain motions, approximating the potential of mean force by a short-ranged harmonic force field is no worse than using such an approximation for the potential energy.

A harmonic protein model consisting of only the C_α positions has been used previously for a theoretical prediction of the temperature factors obtained during crystallographic structure determination.¹⁸ A very good agreement with experiment was observed but required a one-parameter fit for each specific protein. Because temperature factors can also be obtained from normal modes (but only under the assumption that the harmonic approximation is valid even for amplitudes corresponding to physiological temperatures), this result adds support to the suitability of simplified models for the description of slow collective motions.

Analysis of Deformations

Once the normal modes have been calculated, the physically relevant information must be extracted from them. For analyzing domain motion, the most important information is the location of relatively rigid domains and of the more flexible regions between them. Rigid domains are characterized by the absence of local deformations in the low-frequency modes.

A useful measure for the amount of local deformation in continuous media is the energy density due to the deformation as a function of position. A similar quantity can be defined for a point mass system with short-range interactions by distributing each energy term among the atoms involved and summing up the contributions for each atom. This quantity is particularly simple for the pair interaction force field described in the section, Simplified Force Field. The energy contribution for atom i is given by

$$E_i = \frac{1}{2} \sum_{j=1}^N k(\mathbf{R}_{ij}^{(0)}) \frac{|\mathbf{d}_i - \mathbf{d}_j \cdot \mathbf{R}_{ij}^{(0)}|^2}{|\mathbf{R}_{ij}^{(0)}|^2}, \quad (9)$$

where \mathbf{d}_i is the displacement of atom i in the mode to be analyzed and $k(\mathbf{r})$ is given by equation 8. A rigid domain can be identified as a region in which the values of E_i are smaller than a suitably chosen limit.

Care must be taken when comparing deformational energies between different modes or different proteins, because the E_i depend on the amplitude of the displacement. A suitable normalization must be applied to ensure the comparability of the energies. The normalization factor can be deduced from the

condition that the deformation measure for noninteracting identical copies of the system must be equal to the original ones. This leads to the normalized deformation measure

$$D_i = \frac{N}{\sum_{j=1}^N |\mathbf{d}_j|^2} E_i \quad (10)$$

Because the deformation measure is quadratic in the atomic displacement, a meaningful combined deformation measure for several normal mode vectors can be obtained by averaging the values for each individual mode.

The deformation analysis described in this section is not limited to normal modes. With a small modification, it can also be applied to other sets of atomic displacement vectors, e.g., the difference of two structures obtained experimentally. The modification is necessary because such displacement vectors describe finite configurational changes, whereas normal modes are infinitesimal ones. This difference is important in the presence of rotations. A set of displacement vectors describing a global rotation, for example, would contain a combined rotation and deformation when interpreted infinitesimally. Equation 9 must therefore be replaced by its finite displacement analogue

$$E_i = \frac{1}{2} \sum_{j=1}^N k(\mathbf{R}_{ij}^{(0)}) [|\mathbf{R}_{ij}^{(0)} + \mathbf{d}_i - \mathbf{d}_j| - |\mathbf{R}_{ij}^{(0)}|]^2. \quad (11)$$

Description of Domain Motions

Once the domains are identified as sufficiently rigid regions in the protein, their motion in a specified normal mode can be described by a rigid-body motion. The general form of an infinitesimal rigid-body motion is

$$\mathbf{d}_i = \mathbf{T} + \Phi \times \mathbf{R}_i \quad (12)$$

The parameter \mathbf{T} describes the translational contribution and depends on the choice of the coordinate origin. The parameter Φ describes the direction of the rotation axis and the rotational amplitude. Equation 12 describes the absolute motion of a domain relative to a fixed coordinate system; the relative motion between two domains is described by the difference of the parameters \mathbf{T} and Φ obtained for the two domains separately.

It can be shown that any rigid-body motion can be decomposed into rotation around an axis and translation along this axis (see, for example, Goldstein¹⁵). The direction of this axis must obviously be given by Φ ; its position in space is defined by the condition that points on the rotation axis must remain on it during the rigid-body motion. One specific point on

the rotation axis is then given by

$$\mathbf{R}_{\text{axis}} = \frac{\Phi \times \mathbf{T}}{|\Phi|^2}. \quad (13)$$

The vector describing the translational motion along the axis is

$$\mathbf{T}^* = \frac{\mathbf{T} \cdot \Phi}{|\Phi|^2} \Phi. \quad (14)$$

In practice, the domains are not exactly rigid, and the parameters \mathbf{T} and Φ must be obtained by a least-squares fit to equation 12. The calculation of \mathbf{R}_{axis} and \mathbf{T}^* then provides a useful description for visualizing the rigid-body motion by a line that represents at the same time the axis of rotation and the direction of translation.

The exact set of atoms used in the fit should not make an important difference if the domains have been chosen from sufficiently rigid regions of the protein, because the rigid-body parameters for all parts of a rigid regions are obviously the same. Repeating the fit with somewhat different domain definitions is therefore a useful check to verify that the domains have been chosen in a reasonable way.

RESULTS

Accuracy and Performance of the Normal Mode Calculations

To explore the applicability of the methods presented in the previous section, three test systems have been used. The smallest one is crambin, a small protein consisting of 46 amino acids. Such a small protein provides a good “worst case” for methods that are designed for large proteins. The second test system is lysozyme, a well-studied protein with two domains and a characteristic hinge bending motion. The third test system is ATCase, the largest protein to which a standard normal mode analysis has been applied until now.¹⁴

Most calculations were performed on an Hewlett-Packard Vectra VA computer (PentiumPro at 200 MHz, 64 MB of RAM) running the Linux operating system (Hewlett-Packard Company, Palo Alto, CA). Only the full Cartesian normal mode analysis of lysozyme had to be run on a larger machine due to memory requirements; a Hewlett-Packard J282 (PA-8000 processor at 180 Mhz, 512 MB of RAM) was used. The programs are written in Python and C and make use of the Molecular Modeling Toolkit (MMTK)¹⁹ for standard tasks such as system construction, minimization, and normal mode analysis. The Amber 94 force field²⁰ has been used as implemented in MMTK; nonbonded interactions were calculated without cutoff. All proteins were treated in vacuum. Crambin (Protein Data Bank [PDB] entry 1CBN) and turkey egg white lysozyme (PDB

entry 135L) were minimized up to a remaining energy gradient of 10^{-4} kJ/mol/nm using the conjugate gradient minimizer in MMTK. For the Fourier basis normal mode calculations with the Amber 94 force field, the second-derivative matrix for the reduced subspace was evaluated by finite-difference differentiation along the basis vectors to avoid storing the large Cartesian second-derivative matrix. For the short-ranged simplified force field, the Cartesian matrix was stored in the sparse matrix format implemented in MMTK, and the reduced subspace matrix was obtained by multiplication with the basis vectors.

Two sets of normal modes are compared by calculating the overlap matrix, which contains the scalar products of each vector in the first set with each vector in the second set,

$$O_{ij} = \mathbf{v}_i^* \cdot \mathbf{w}_j^*, \quad (15)$$

where \mathbf{v}_i^* and \mathbf{w}_j^* are the two sets of mass-weighted mode vectors. If the two sets are identical, the result is a unit matrix. For two similar but not identical sets, there will be large values on and close to the diagonal and small values elsewhere. To compare a set of atomic normal mode vectors with a set of residue-based normal mode vectors, each atomic mode vector is transformed into a residue vector by calculating the residue center-of-mass displacements from the atomic displacements. The resulting set of displacement vectors is no longer strictly orthonormal, but a meaningful comparison requires only one of the two vector sets to be orthonormal.

Because the graphical representation of such a matrix is not always easy to interpret, it is useful to define a simpler one-dimensional measure of similarity. For two full sets of modes, the squares of the scalar products add up to one along each direction. They can therefore be considered a kind of “distribution,” and the width of this distribution indicates over how many modes of the other set any given mode is spread. The exact definition of the spread is

$$s_i = \sqrt{\sum_j J^2 O_{ij}^2 - \left(\sum_j j O_{ij}^2 \right)^2}, \quad (16)$$

in analogy to the definition of the standard deviation of a probability distribution. When not all modes of the set labelled by j are available, the overlaps in equation 16 must be scaled to ensure that $\sum_j O_{ij}^2 = 1$.

The spread indicates how many modes in one set have a significant overlap with a specific mode in the other set. For two identical sets of modes, the overlaps O_{ij} are nonzero only for $i = j$, and the spread becomes zero. For two totally unrelated sets of orthogonal displacement vectors, all the overlaps O_{ij} are of the same order of magnitude, and the spread grows with the number of modes. However, care should be taken not to misinterpret this quantity.

The “distributions” are not statistical distributions, much less the commonly assumed Gaussian ones. A large spread can indicate either many significantly non-zero overlaps for a given mode or few non-zero overlaps that are widely spaced. In both cases one would speak of a bad agreement between the two mode sets. Conversely, a small spread means that there are few significant overlaps *and* that they correspond to neighboring modes. This shows that the spread is indeed a useful measure of similarity.

Small proteins: crambin

Crambin is a very small protein, consisting of a single chain of 46 amino acids with a total of 642 atoms. It is too small to have recognizable domains or domain motions; therefore, it can be considered a “worst case” for the methods presented in this article, which were designed for large proteins. Crambin is small enough to permit a full Cartesian normal mode analysis, which took 40 minutes of CPU time. Fourier basis normal mode analyses were performed for various values of λ_{\min} ; the smallest for $\lambda_{\min} = 1.6$ nm used 96 mode vectors and took 3 minutes. For comparison, a dihedral subspace basis allowing variation of only the ϕ - and ψ -angles was also used; this basis consists of 100 vectors and is thus comparable in size to the smallest Fourier basis.

Figure 1A shows an example of a full overlap matrix, which compares the full Cartesian modes to the Fourier basis modes at $\lambda_{\min} = 1.2$ nm, which is a set of 246 modes. It is clear that the overlap matrix is dominantly diagonal, but there are also significant overlap values somewhat away from the diagonal. Figure 1B shows the overlap for the first three Cartesian modes in detail. A better quantitative comparison can be obtained from Figure 2, which shows the spread of each exact mode over the Fourier basis modes, as defined in equation 16. Figure 2A compares two Fourier basis mode sets of different size and the $\phi - \psi$ -angle basis modes. The spread for all three sets is surprisingly similar; one would expect a much smaller spread for a basis of 246 modes than for one of 96 modes. However, the spread must decrease significantly as the number of basis vectors grows, because it is zero for a full basis (1926 modes in the case of crambin). Figure 2B shows that this happens indeed, but suddenly around a rather low value of $\lambda_{\min} \approx 0.55$ nm.

For another comparison, the full Cartesian modes were calculated with the simplified force field described in the section, Simplified Force Field; that calculation required 27 minutes. Another set of modes with the same force field and a Fourier basis with $\lambda_{\min} = 1.6$ nm was obtained in 30 seconds. Figure 2C shows the spread for these two mode sets. There is little difference compared with the 1.6 nm Fourier basis mode set obtained with the Amber 94 force field.

An explanation for the observed dependence of the spread on the various approximations can be obtained from the frequency spectrum of crambin, shown in Figure 3. The most striking feature of this spectrum (which is typical for proteins in general, because the low-frequency modes that are characteristic for a specific protein occupy only a very small interval close to zero) is the absence of any modes in the frequency interval from 55 THz to 85 THz (1800 cm^{-1} to 2800 cm^{-1}). An analysis of the modes corresponding to the two well-separated blocks shows that the high-frequency modes are bond stretching modes involving hydrogen atoms, whereas all other modes fall into the lower frequency block, with bond angle modes at the upper end of the spectrum. Bond angle vibrations involve the relative motion of atoms at distances of 0.15 to 0.25 nm, which would be covered exactly by Fourier bases with $\lambda_{\min} \approx 0.4$ nm or less. The spread thus decreases sharply as soon as the bases with decreasing λ_{\min} start to describe bond angle vibrations in detail. The bond stretching modes are separated well enough in the frequency spectrum that they can be considered independent of the other motions. A similar observation is well known from molecular dynamics simulations: bond stretching modes can be eliminated by constraints without changing the dynamics of the remaining degrees of freedom significantly, but eliminating the bond angle movements as well causes important modifications.²¹ However, this is no longer true for the simplified force field, whose frequency spectrum (scaled to make the highest frequencies of both spectra equal) is also shown in Figure 3. This force field is not sufficiently detailed to describe the dynamics of a protein reasonably well at such small length and time scales.

From a practical point of view, the results presented in this section show that if one is willing to accept the accuracy of normal modes shown in Figure 2A, then a small basis and a simplified force field will yield an answer in a small fraction of the time required for a full normal mode calculation. A significantly better normal mode analysis requires a computational effort that is close to that of a full Cartesian calculation. It must also be kept in mind that the Amber 94 force field used for the “exact” calculation was not designed specifically for normal mode analysis. Although it is more detailed than the simplified force field and can therefore be expected to yield normal modes that are closer to reality, there is no way to quantify the expected accuracy of such modes. There may therefore be no justification for interpreting normal mode calculations beyond the level of agreement established by Figure 2.

Medium size proteins: lysozyme

Lysozyme is a popular test case for studying domain motions. It has two relatively rigid domains connected by a more flexible hinge region. The hinge

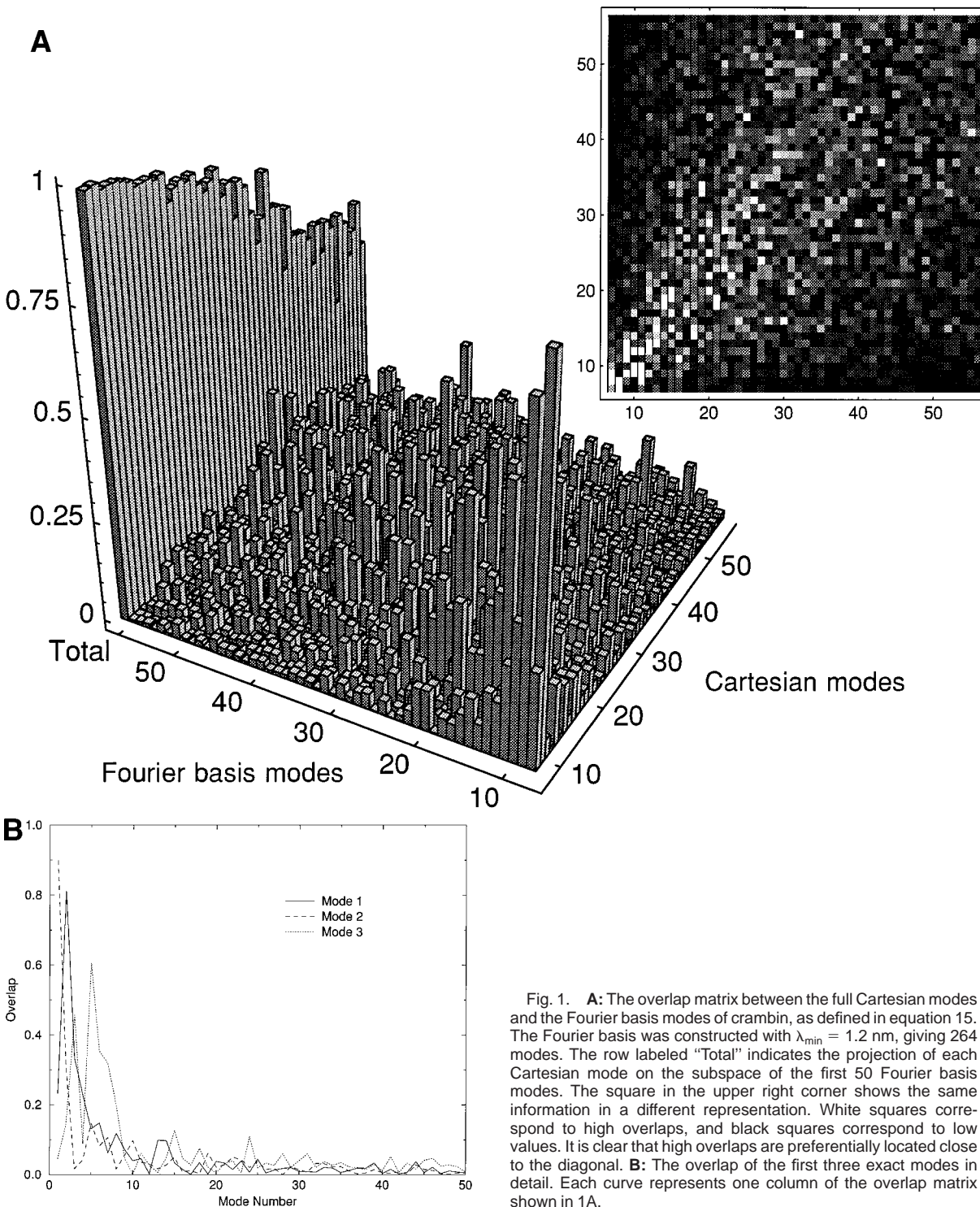


Fig. 1. **A:** The overlap matrix between the full Cartesian modes and the Fourier basis modes of crambin, as defined in equation 15. The Fourier basis was constructed with $\lambda_{\min} = 1.2$ nm, giving 264 modes. The row labeled "Total" indicates the projection of each Cartesian mode on the subspace of the first 50 Fourier basis modes. The square in the upper right corner shows the same information in a different representation. White squares correspond to high overlaps, and black squares correspond to low values. It is clear that high overlaps are preferentially located close to the diagonal. **B:** The overlap of the first three exact modes in detail. Each curve represents one column of the overlap matrix shown in 1A.

motion has been studied by both normal mode analysis and molecular dynamics.⁶ With 129 amino acids and 1950 atoms, lysozyme is just small enough to permit a full Cartesian normal mode analysis on a well-equipped workstation.

Due to memory requirements, the full normal mode calculation for lysozyme had to be done on a

different machine that is approximately three times as fast as the machine used for the other calculations; on that machine it took 4.6 hours. For comparison, a Fourier basis with $\lambda_{\min} = 1.2$ nm yields 564 modes in 2.3 hours on the slower machine. With the simplified force field shown in this article and the same basis, the calculation could be completed in 22

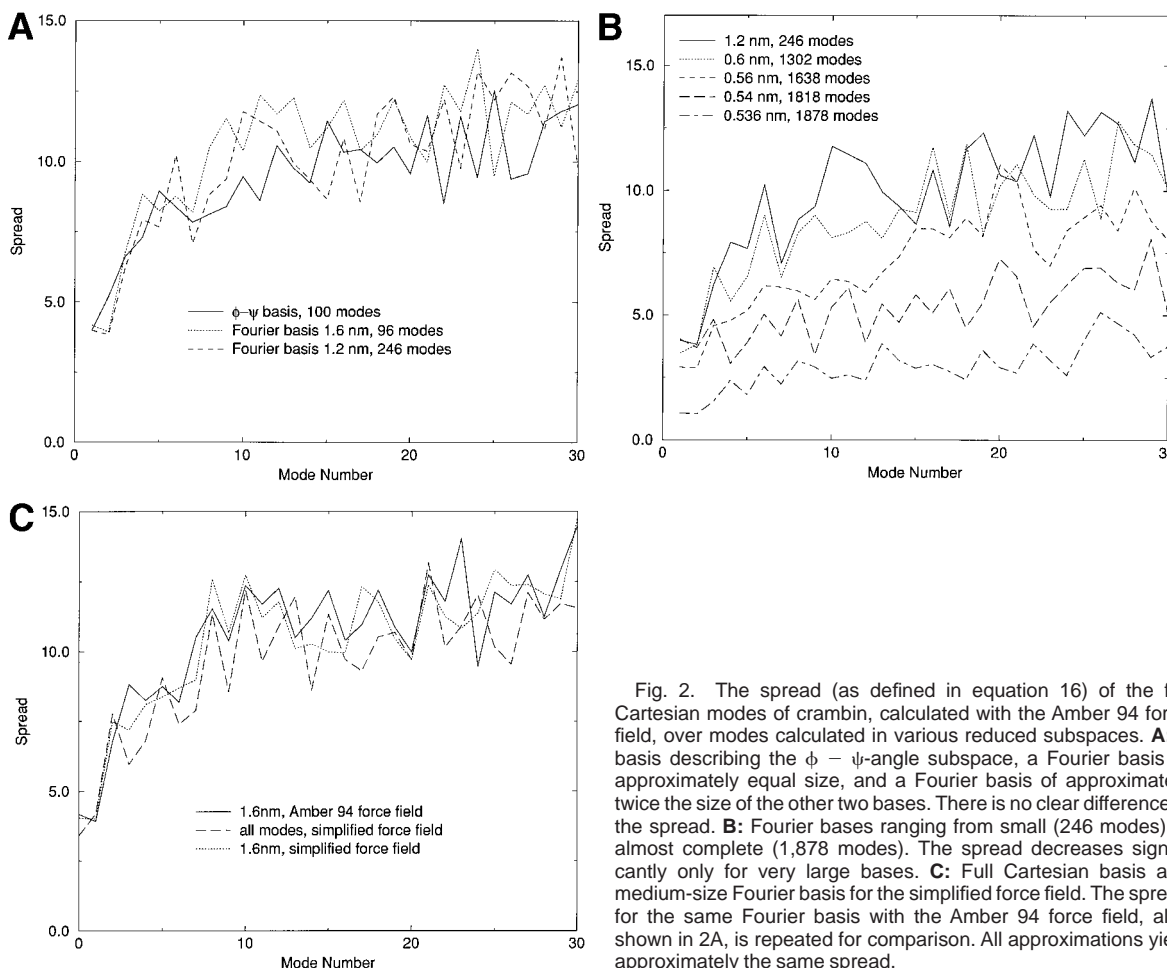


Fig. 2. The spread (as defined in equation 16) of the full Cartesian modes of crambin, calculated with the Amber 94 force field, over modes calculated in various reduced subspaces. **A:** A basis describing the $\phi - \psi$ -angle subspace, a Fourier basis of approximately equal size, and a Fourier basis of approximately twice the size of the other two bases. There is no clear difference in the spread. **B:** Fourier bases ranging from small (246 modes) to almost complete (1,878 modes). The spread decreases significantly only for very large bases. **C:** Full Cartesian basis and medium-size Fourier basis for the simplified force field. The spread for the same Fourier basis with the Amber 94 force field, also shown in 2A, is repeated for comparison. All approximations yield approximately the same spread.

minutes. Using the simplified protein model shown in this article and a Fourier basis with the same cutoff (resulting in only 402 modes due to the smaller model), the normal mode calculation took only 52 seconds. In addition, the last two calculations could have been done without a prior lengthy energy minimization.

The spread for all approximations (Fig. 4) is of similar size and also close to the values for crambin. Even the rather drastic model simplification from 1950 atoms to 129 point masses representing the residues does not lead to a degradation of the similarity of the low-frequency modes. It can be concluded that the characterization of low-frequency modes does not require a detailed description of the interactions, but can be considered to be essentially a structural property.

Large proteins: ATCase

ATCase is a large allosteric protein (2,760 residues) that exhibits large rearrangements of essentially rigid domains during the allosteric transition.²² The first 53 normal modes of ATCase have been calculated by Thomas et al.¹⁴ using a matrix

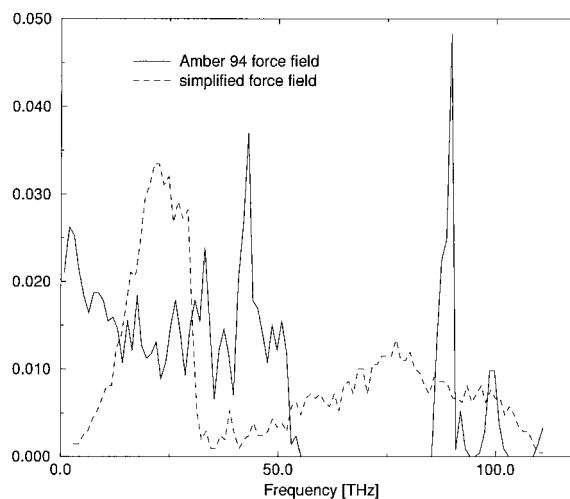


Fig. 3. The density of vibrational frequencies for crambin, calculated with the Amber 94 force field and the simplified force field. The frequency spectrum of the simplified force field is simpler and does not show the clear separation between hydrogen bond stretching modes and all other modes that is characteristic of the spectrum obtained with the Amber 94 force field.

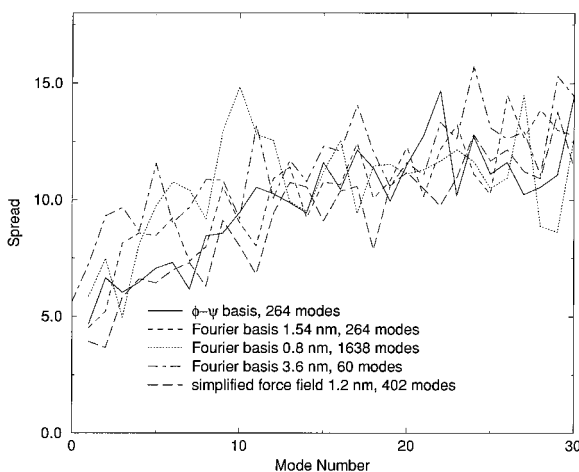


Fig. 4. The spread of the full Cartesian modes of lysozyme over modes calculated in various approximations. All approximations show approximately equal spread, which is even approximately the same as the spreads for crambin shown in Figure 2.

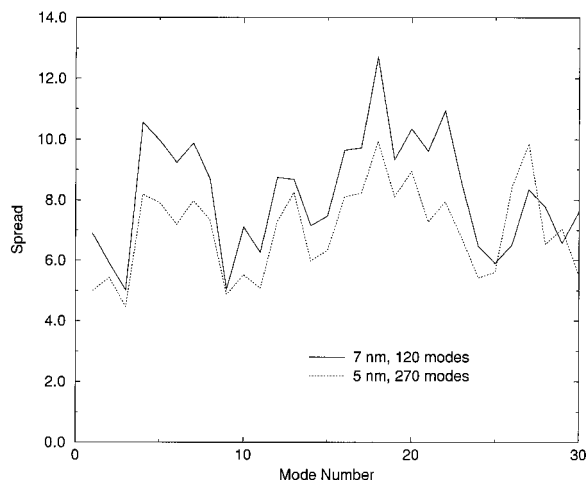


Fig. 5. The spread of the full Cartesian modes of ATCase over modes calculated with the simplified protein model described in this article. Again, the spread does not depend strongly on the approximation level and is of the same order of magnitude as for crambin and lysozyme.

partitioning method¹³; this calculation required 690 hours on a Cray C98 supercomputer. Using the current simplified protein model with the point masses located at the centers of mass of the residues, and a Fourier basis with $\lambda_{\min} = 5$ nm, 270 low-frequency normal modes were obtained in 9 minutes. A smaller set of 120 modes was calculated by choosing $\lambda_{\min} = 7$ nm in 3 minutes. The same T state configuration as in Thomas et al.¹⁴ was used to ensure comparability. It describes a partial united-atom model (only the polar hydrogen atoms are represented explicitly) that was minimized with the CHARMM force field.

Although ATCase is much larger than crambin and lysozyme, the spread shown in Figure 5 is very similar but in general somewhat smaller. Again the influence of the size of the basis is weak, although the larger basis yields a smaller spread for almost all modes. The extremely similar behavior of the spread over a wide range of system sizes permits the hypothesis that approximate normal mode calculations have a universal accuracy that depends little on system size and detail of the models. Because the so-called “exact” model (classical point masses with an empirical force field) is itself an approximation whose accuracy for normal mode calculations is essentially unknown, it is questionable whether the results of any normal mode analysis can be interpreted beyond the level of precision at which all the models tested here are equivalent.

Domain Analysis

The deformation analysis described in the section, Analysis of Deformations, has been applied to the two larger test cases, lysozyme and ATCase. The results are shown in Figure 6. The colors indicate the

deformation, with a color scale ranging from blue (small deformation) via green and yellow to red (high deformation). The color scales for the two proteins were derived independently and should not be compared. For lysozyme, the deformation was calculated as the sum of the contributions of the first four modes (obtained with the Amber 94 force field and a Fourier basis), because only those modes showed predominantly interdomain motion. For ATCase the first 15 modes (obtained with a 5 nm Fourier basis and the simplified protein model shown in this article) were used. In both cases, a small variation in the number of modes does not lead to a visible difference.

For lysozyme, it is immediately clear that there are at least two domains, a small one at the left (with residues 47–49 and 68–70 at the core) and a large one at the right (defined by residues 12, 14, 22, 28, 112, and 117). These two domains are connected by a more flexible region, with the most strongly deformed region close to the active site, between residues 44 and 109. The atomic displacements due to the first normal mode are indicated in the figure by a vector field representation. The first three modes describe essentially rotations between the two domains. The rotation axes corresponding to these motions were obtained as described in Description of Domain Motions (using the most rigid 60% of all residues to define the domains) and are also shown in Figure 6; the translational part of the movement turns out to be negligible. Because the rotation axes are almost perpendicular and almost intersect in a single point (which lies within 4 Å of the C_{α} atoms of residues 53, 54, and 58), the three first modes essentially permit free rotation around this point. This seems to contradict another recent study by

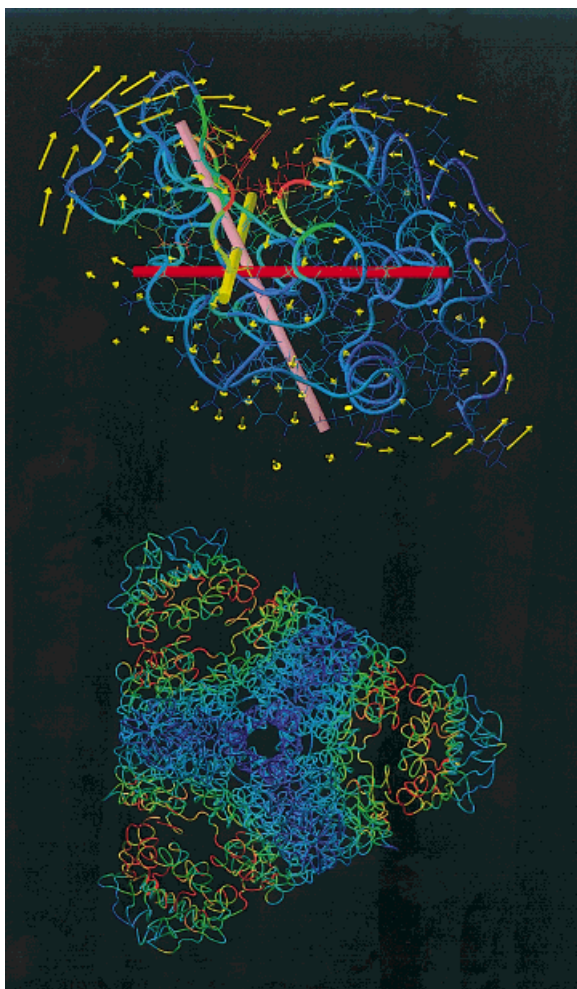


Fig. 6. Deformation in lysozyme (**top**) and ATCase (**bottom**). The colors of the atoms and the backbone indicate the amount of deformation in the first four (lysozyme) or six (ATCase) normal modes; dark blue regions are the most rigid, whereas red regions are strongly deformed. The color scale is different for the two proteins. For lysozyme, the atomic displacement field corresponding to the first mode is indicated by the yellow arrows. The interdomain rotation axes of the first three modes are represented by the colored cylinders in the order yellow, red, magenta. The image was created using the MMTK toolkit,¹⁹ the visualization program VMD,²³ and the rendering program POV-Ray.

Hayward et al.,⁷ which finds only two rotational degrees of freedom. These authors do not distinguish between rigid and more flexible regions and instead attempt to divide the whole protein cleanly into two domains. It is to be expected that such an approach leads to fewer modes describing interdomain motion. The intersection point of the two rotation axes obtained by Hayward et al. lies sufficiently close to the one found here, such that the two analyses can be considered to be essentially in agreement.

A similar picture produced from the full Cartesian normal modes (not shown) differs essentially by a smaller highly flexible region around the active site.

This means that the transition region between the domains is smaller. Because the Fourier basis excludes sharp transitions between domains, this is not surprising. However, the two domains are the same. There are also three modes describing interdomain rotations, and their axes are similarly perpendicular and intersect in one point, which is almost identical to the one found from the Fourier basis modes. However, the directions of the axes are different, indicating that the three modes describing interdomain motions, which are very close in frequency, are different linear combinations than for the approximate mode set. Ultimately, the conclusions that can be drawn with any certainty from the two mode sets are the same.

ATCase, being a much larger protein, shows a more complicated domain structure. It is composed of six regulatory chains, arranged in three dimers that form the tips of the triangle, and six catalytic chains, arranged in two dimers that form the core of the molecule.²² Most of the bottom trimer is covered by the top trimer in Figure 6. Structurally each regulatory chain can be divided into an allosteric and a zinc domain that are linked by a very flexible loop. The catalytic chains can be divided into an aspartate domain and a carbamyl phosphate domain; because these domains are linked tightly by two helices, the division is less evident. Figure 6 shows one large rigid domain at the core of each trimer and a much smaller rigid domain in the exterior parts of each dimer. This is in agreement with a detailed analysis of the first 53 full Cartesian modes by Thomas et al.¹⁴ In fact, a deformation analysis based on these modes (not shown) does not show any clear difference from Figure 6. A more detailed description of the domains and domain motions in ATCase requires more sophisticated analysis techniques than the simple rigid-body fit used for lysozyme, because there are more domains and more different motions for each domain. Moreover, the domains split into recognizable subdomains in some higher modes. Suitable techniques for the identification of such a domain hierarchy and the description of its motions will be presented in a separate article.

CONCLUSION

The main goal of this article has been the presentation and demonstration of several new methods for analyzing domain motion in proteins. A general subspace for calculating low-frequency modes has been developed, and a simplified force field and protein model have permitted a drastic reduction of the computational resources that are required for a normal mode analysis. Furthermore, a general technique for detecting rigid domains has been presented. Together these techniques transform normal-mode-based dynamic domain analysis from a costly technique reserved for important cases into a routine technique that is easily applied by experimentalists

and theoreticians to gain a first insight into the low-frequency dynamics of a protein. An implementation of these techniques in the form of a ready-to-use program is available from the author.

In addition to obvious practical benefits, the new methods also provide new physical insight into the nature of the potential energy surface of proteins and the significance of normal mode calculations. The fact that a drastically simplified force field, which does not take into account fundamental properties such as atom type or bond structure, can correctly identify low-frequency modes shows that the distinction between low- and high-frequency motion is essentially a structural property independent of the details of atomic interactions.

This independence has several important implications for normal mode analysis. With standard force fields, the existence of many distinct local energy minima raises the question whether modes calculated for any one minimum can be considered typical for all nearby minima. The simplified force field varies smoothly with changes of the input configuration, implying a smooth continuous change of the normal modes. Because the simplified force field and the Amber 94 and CHARMM 19 force fields lead to normal modes that agree within a well-defined and seemingly universal accuracy, it follows that the variation of normal modes between nearby local minima must stay within the same limits.

Moreover, the fact that the simplified protein model is able to reproduce the low-frequency modes of large proteins rather well explains why normal mode analysis, despite its exploration of only a single local energy minimum of the configurational space of the system, can make meaningful predictions for the system in its real physiological environment. Such environments have temperatures at which entropic effects are not negligible, and hence the relevance of studying minima of potential energy is questionable. Instead, the free energy as a function of slow variables should be analyzed. As explained in this article, the simplified protein model can in fact be regarded as a crude approximation to the free energy as a function of residue positions. Because such a model produces essentially the same low-frequency motions as an atomic model with a potential energy surface, it can be concluded that the neglect of entropic effects in standard normal mode analysis has no important consequences as far as domain motions are concerned.

The implication of these observations for the energy landscape of proteins is that the multiple local minima of the potential energy in the subspace of low-frequency motions and the corresponding smoothed-out minima of the free energy profile must have similar shape. This shape is essentially determined by the condition that deformations should be limited to small regions and/or regions with a low

atom density, because a low atom density implies a lower energetic cost of deformations.

The comparison of normal modes obtained with various models and reduced subspaces has shown that there is a seemingly universal level of precision up to which all calculations produce the same result. Agreement to a much greater precision could only be obtained for almost identical descriptions. Because none of the models ever used for normal mode analysis of proteins can be claimed to be exact or even significantly better than other models, the practical conclusion is that no normal mode analysis should be interpreted beyond the precision at which all models yield the same results. However, this level of precision is sufficient to identify domains and their large-scale motions.

ACKNOWLEDGMENTS

The author thanks Dr. M.J. Field and Dr. A. Thomas for helpful discussions and provision of the reference normal mode vectors for ATCase from Thomas et al.¹⁴

REFERENCES

1. Case, D.A. Normal mode analysis of protein dynamics. *Curr. Opin. Struct. Biol.* 4:285–290, 1994.
2. Go, N., Noguti, T., Nishikawa, T. Dynamics of a small globular protein in terms of low-frequency vibrational modes. *Proc. Natl. Acad. Sci. USA* 80:3696–3700, 1983.
3. Amadei, A., Linssen, A.B.M., Berendsen, H.J.C. Essential dynamics of proteins. *Proteins* 17:412–425, 1993.
4. Kottalam, J., Case, D.A. Langevin modes of macromolecules: Application to Crambin and DNA hexamers. *Biopolymers* 29:1409–1421, 1990.
5. Hayward, S., Kitao, A., Hirata, F., Go, N. Effect of solvent on collective motions in globular proteins. *J. Mol. Biol.* 234:1207–1217, 1993.
6. Horiuchi, T., Go, N. Projection of Monte-Carlo and molecular dynamics trajectories onto the normal mode axes: Human lysozyme. *Proteins* 10:106–116, 1991.
7. Hayward, S., Kitao, A., Berendsen, H.J.C. Model-free methods of analyzing domain motions in proteins from simulations: A comparison of normal mode analysis and molecular dynamics simulation. *Proteins* 27:425–437, 1997.
8. Elber, R., Karplus, M. Multiple conformational states of proteins: A molecular dynamics analysis of myoglobin. *Science* 235:318–321, 1987.
9. Frauenfelder, H., Parak, F., Young, R.D. Conformational substates in proteins. *Ann. Rev. Biophys. Biophys. Chem.* 17:451–479, 1988.
10. Lamy, A.V., Souaille, M., Smith, J.C. Simulation evidence for experimentally detectable low-temperature vibrational inhomogeneity in a globular protein. *Biopolymers* 39:471–478, 1996.
11. Brooks, B.R., Janežič, D., Karplus, M. Harmonic analysis of large systems. I. Methodology. *J. Comp. Chem.* 16:1522–1542, 1995.
12. Hao, M.H., Scheraga, H.A. Analyzing the normal mode dynamics of macromolecules by the component synthesis method: Residue clustering and multiple-component approach. *Biopolymers* 34:321–334, 1994.
13. Mouawad, L., Perahia, D. Diagonalization in a mixed basis: A method to compute low-frequency normal modes for large macromolecules. *Biopolymers* 33:599–611, 1993.
14. Thomas, A., Field, M.J., Mouawad, L., Perahia, D. Analysis of the low frequency normal modes of the T-state of aspartate transcarbamylase. *J. Mol. Biol.* 257:1070–1087, 1996.

15. Goldstein, H. "Classical Mechanics." Reading, MA: Addison-Wesley, 1980.
16. Teeter, M.M., Case, D.A. Harmonic and quasiharmonic descriptions of crambin. *J. Phys. Chem.* 94:8091–8097, 1990.
17. Tirion, M.M. Low-amplitude elastic motions in proteins from a single-parameter atomic analysis. *Phys. Rev. Lett.* 77:1905–1908, 1996.
18. Bahar, I., Atilgan, A.R., Erman, B. Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential. *Folding Design* 2:173–181, 1997.
19. 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>
20. Cornell, W.D., Cieplak, P., Bayly, C.I., et al. A second generation force field for the simulation of proteins and nucleic acids. *J. Am. Chem. Soc.* 117:5179–5197, 1995.
21. van Gunsteren, W.F., Karplus, M. Effect of constraints on the dynamics of macromolecules. *Macromolecules* 15:1528–1544, 1982.
22. Lipscomb, W.N. Aspartate transcarbamylase from *Escherichia coli*: Activity and regulation. *Adv. Enzymol. Relat. Areas Mol. Biol.* 68:67–151, 1994.
23. Humphrey, W., Dalke, A., Schulten, K. VMD—Visual Molecular Dynamics. *J. Mol. Graph.* 14:33–38, 1996.