JCTC Journal of Chemical Theory and Computation

Harmonic Vibrational Analysis in Delocalized Internal Coordinates

Frank Jensen* and David S. Palmer†

Department of Chemistry, University of Aarhus, DK-8000 Aarhus, Denmark

Received August 18, 2010

Abstract: It is shown that a principal component analysis of a large set of internal coordinates can be used to define a nonredundant set of delocalized internal coordinates suitable for the calculation of harmonic vibrational normal modes. The selection of internal coordinates and the principal component analysis provide large degrees of freedom in extracting a nonredundant set of coordinates, and thus influence how the vibrational normal modes are described. It is shown that long-range coordinates may be especially suitable for describing low-frequency global deformation modes in proteins.

Introduction

The nuclear motion of atoms in a molecule can, within the Born−Oppenheimer approximation, be determined by solving the nuclear Schrödinger equation, where the solution to the electronic Schrödinger equation plays the role of a potential energy surface (PES). Near a minimum on the PES, the potential energy can be approximated by a second-order Taylor expansion, and solutions to the nuclear Schrödinger equation can be expressed as a product of solutions to the one-dimensional harmonic oscillator problem. If the quantum aspects of the nuclei are ignored, the motion of the nuclei can be determined by solving the corresponding classical vibrational equation. Harmonic vibrational analyses are, in many cases, of sufficient accuracy for identifying unknown species by comparing calculated vibrational spectra with experimentally observed ones and also form the starting point for more advanced treatment of anharmonic vibrations.[1]

The eigenvalues from the solution of the harmonic vibrational equation are proportional to the vibrational frequencies, and the associated eigenvectors are the normal coordinates, often called normal modes. The vibrational frequencies contain information regarding the curvature of the underlying PES, and it is commonly assumed that the directions of the normal modes associated with the lowest vibrational frequencies can be used to search for low-energy transition states connecting to other low-energy minima on the PES.[2] When applied to biomolecules, it has been observed that low-frequency normal modes in many cases can be used to rationalize even large-scale domain movements in proteins.[3−9] More recently, such low-frequency normal modes have been used to bias molecular dynamics simulations, with the purpose of simulating conformational transitions that otherwise would be too slow to be computationally feasible.[10] We note in passing that many of these conclusions are based on normal modes calculated by elastic network models,[9] rather than atomistic force fields, but there is substantial evidence that the two types of normal modes are in reasonably good agreement with each other.[11]

While the (tangential) direction of a normal mode at a stationary point is independent of the coordinates used for expressing the vibrational problem, the atomic coordinates for any finite displacement along a normal mode will depend on the coordinates used for solving the vibrational equation.[12,13] Cartesian coordinates have been used almost exclusively, since the vibrational problem in this case can be formulated as a simple diagonalization of a mass-weighted force constant matrix. When applied to systems with hundreds or thousands of atoms, low-frequency modes often involve a large fraction of all of the coordinates, and the degree of localization/delocalization of a vibrational mode can be quantified in terms of a localization or collectivity index.[14] The vibrational problem, however, can be solved in any nonredundant set of coordinates,[15] and in the present paper, we explore whether it is possible to select other sets of coordinates that allow a more concise description of, in particular, the low-frequency vibrations in biomolecules like proteins. For small

* Corresponding author. E-mail: frj@chem.au.dk.

† Current address: Max Planck Institute for Mathematics in the Sciences, Inselstrasse 22, DE-04103 Leipzig, Germany.

10.1021/ct100463a
Published on Web 12/03/2010

systems, there is a substantial amount of work reported for quantum vibrational analyses in internal coordinates,[16] while the applications to larger systems are significantly more scarce. Kamiya et al. have discussed how to efficiently calculate the second derivative matrix in internal coordinates from a parametrized energy function and use this to perform vibrational analysis in selected sets of internal coordinates.[17] Transformation from Cartesian coordinates to internal coordinates for analysis purposes has also been discussed.[18,19]

Theory

The determination of vibrational normal coordinates from a set of general coordinates goes back to the seminal work of Wilson et al.,[20,21] where a set of $3N - 6$ nonredundant internal coordinates were assumed to be available. Fogarasi et al. introduced the use of natural internal coordinates for geometry optimization,[22] and several other groups have generalized this idea to extract $3N - 6$ nonredundant linear combinations, often denoted delocalized internal coordinates,[23] from a large set of redundant primitive internal coordinates.[24−28] We will in the present paper use these concepts to determine vibrational normal coordinates as a linear combination of delocalized internal coordinates extracted from a (large) set of primitive internal coordinates. Lower case bold symbols like **a** will, in the following, denote a vector with a length given by the number of variables, while a capital symbol like **A** denotes a matrix containing all **a** vectors as columns.

We will assume a nonlinear system with N atoms and $3N$ Cartesian coordinates contained in a vector **x**, and thus $3N - 6$ vibrational degrees of freedom. In Cartesian coordinates, the classical kinetic and potential energy operators can be written as in eqs 1−3,[29] where the harmonic approximation has been employed for the potential energy and dotted vectors indicate time derivatives.

$$2\mathbf{T} = \sum_i^N \dot{\mathbf{x}}_i m_i \dot{\mathbf{x}}_i = \dot{\mathbf{x}}^t \mathbf{M} \dot{\mathbf{x}} \tag{1}$$

$$2\mathbf{V} = \sum_{ij}^N \Delta\mathbf{x}_i F_{ij} \Delta\mathbf{x}_j = \Delta\mathbf{x}^t \mathbf{F} \Delta\mathbf{x} \tag{2}$$

$$F_{ij} = \frac{\partial^2 E}{\partial x_i \partial x_j};\ M_{ij} = \delta_{ij} M_i \tag{3}$$

The system can alternatively be described by a set of K internal coordinates **q**, where $K \geq 3N - 6$. The internal coordinates will typically be bond stretch coordinates (distance between two atoms), bending coordinates (angle between three atoms), and torsional coordinates (dihedral angle defined by four atoms) but may also be other types of coordinates.[30] A set of all possible stretch (S), bend (B), and torsion (T) coordinates can be generated automatically from a set of Cartesian coordinates by employing suitable atomic radii for determining the bonding pattern,[27,28] and such a set will in the following be denoted the primitive coordinates.

Each internal coordinate q_i is a function of the Cartesian coordinates and can be Taylor expanded as shown in eq 4.

$$q_i = q_i(\mathbf{x}) = q_i(\mathbf{x}_0) + \frac{\partial q_i}{\partial x}(\mathbf{x} - \mathbf{x}_0) + \ldots \tag{4}$$

At a stationary point, only the first derivative of the internal coordinates with respect to Cartesian coordinates is required for transforming the kinetic and potential energy operators to internal coordinates,[12,31,32] which can be written in terms of the Wilson **B** matrix and the corresponding (generalized) inverse $\mathbf{B}^{-1}$ matrix.

$$\Delta\mathbf{q} = \mathbf{B}\Delta\mathbf{x} \tag{5}$$

$$B_{ij} = \frac{\partial q_i}{\partial x_j} \tag{6}$$

$$\Delta\mathbf{x} = \mathbf{B}^{-1}\Delta\mathbf{q} \tag{7}$$

$$\mathbf{B}^{-1} = \mathbf{B}^t(\mathbf{B}\mathbf{B}^t)^{-1} = \mathbf{B}^t\mathbf{G}^{-1} \tag{8}$$

The vibrational part of the **B** matrix is rectangular with dimension $(3N - 6) \times 3N$, where the remaining six vectors describing overall translation and rotation can be determined by the Eckart conditions.[33] The kinetic and potential energies transformed to internal coordinates are shown in eqs 9−11.

$$2\mathbf{T} = \dot{\mathbf{q}}^t((\mathbf{B}^{-1})^t\mathbf{M}\mathbf{B}^{-1})\dot{\mathbf{q}} = \dot{\mathbf{q}}^t\mathbf{G}_\mathbf{m}^{-1}\dot{\mathbf{q}} \tag{9}$$

$$\mathbf{G}_\mathbf{m} = \mathbf{B}\mathbf{M}^{-1}\mathbf{B}^t \tag{10}$$

$$2\mathbf{V} = \Delta\mathbf{q}^t((\mathbf{B}^{-1})^t\mathbf{F}\mathbf{B}^{-1})\Delta\mathbf{q} = \Delta\mathbf{q}^t\mathbf{F}_\mathbf{q}\Delta\mathbf{q} \tag{11}$$

Determining the vibrational normal coordinates corresponds to choosing a linear transformation **L** that simultaneously makes $\mathbf{G_m}^{-1}$ a unit matrix and $\mathbf{F_q}$ a diagonal matrix, which is equivalent to finding the eigenvectors of the $\mathbf{G_m F_q}$ matrix product.[20,21]

$$\mathbf{G_m F_q L} = \mathbf{L}\boldsymbol{\Lambda} \tag{12}$$

Since the $\mathbf{G_m}$ matrix is nondiagonal, the $\mathbf{G_m F_q}$ matrix is nonsymmetric, but the problem can be symmetrized by transforming the **l** coordinates by $\mathbf{G_m}^{-1/2}$.[34]

$$(\mathbf{G_m}^{1/2}\mathbf{F_q}\mathbf{G_m}^{1/2})(\mathbf{G_m}^{-1/2}\mathbf{L}) = (\mathbf{G_m}^{-1/2}\mathbf{L})\boldsymbol{\Lambda} \tag{13}$$

$$(\boldsymbol{G}_\mathrm{m}^{1/2}\boldsymbol{F}_\mathrm{q}\boldsymbol{G}_\mathrm{m}^{1/2})\boldsymbol{U} = \boldsymbol{U}\boldsymbol{\Lambda} \tag{14}$$

$$\boldsymbol{L} = \boldsymbol{G}_\mathrm{m}^{1/2}\boldsymbol{U} \tag{15}$$

The vibrational normal coordinates are defined by the eigenvectors obtained by diagonalization of the $\mathbf{G_m}^{1/2}\mathbf{F_q}\mathbf{G_m}^{1/2}$ matrix, where the square root of the eigenvalues λ is proportional to the vibrational frequencies.[35] The normal coordinates defined by the **u** vectors can be back-transformed to **l** coordinates by multiplying by $\mathbf{G_m}^{1/2}$, where they are given as a linear combination of internal coordinates **q**.

$$\mathbf{l_k} = \sum_i^\mathrm{K} c_{ki} q_i \tag{16}$$

The square of the coefficients c_{ki} determines the contribution of the internal coordinate q_i to the (normalized) $\mathbf{l_k}$ normal

mode. The normal modes can be further transformed to Cartesian coordinates by eq 7, which must be done iteratively as the $\mathbf{B}^{-1}$ matrix only describes the first-order change.[36] We note that the normal modes are orthogonal in the **u** space, but not in the **l** or **x** spaces.

For small systems ($N < \sim 100$ atoms), it is usually possible to select a nonredundant set ($K = 3N - 6$) of primitive coordinates manually. The requirement that the **q** coordinates are nonredundant is equivalent to requiring that the $\mathbf{G_m}$ matrix is nonsingular. In practical applications, a near-singular $\mathbf{G_m}$ must also be avoided for numerical reasons, and a heuristic criterion in terms of its determinant shown in eq 17 has been proposed.[28]

$$-\log(|\mathbf{G_m}|) < \sim(3N - 6) \tag{17}$$

It should be noted that this criterion is rather conservative, as satisfactory numerical accuracy can be obtained for substantially smaller values of the $\mathbf{G_m}$ determinant; for example, **q** coordinates corresponding to $-\log(|\mathbf{G_m}|) \sim 3(3N - 6)$ can be handled without a significant loss of numerical accuracy. The determinant of $\mathbf{G_m}$ is closely related to the condition number, which has been used by Németh et al. as a criterion for selecting primitive internal coordinates in connection with geometry optimization.[37]

For systems with more than ~100 atoms, it is often difficult to define a set of primitive coordinates that does not lead to a near singular $\mathbf{G_m}$ matrix. A redundant set of internal coordinates ($K > 3N - 6$), however, may be transformed into $3N - 6$ linearly independent and $K - (3N - 6)$ dependent coordinates by a transformation matrix **S** (assuming that the primitive coordinates span the full $3N - 6$ vibrational space). The **S** matrix can be chosen to reflect the molecular symmetry (if any) and is therefore often called a symmetry matrix. The **S** matrix has dimension $K \times K$, but only the $3N - 6$ linearly independent coordinates are of interest; i.e., only the $(3N - 6) \times K$ elements of the **S** matrix are significant. We will denote the symmetry coordinates by **d**, and the transformation can be written as in eq 18.

$$\mathbf{d} = \mathbf{S}^t\mathbf{q} \tag{18}$$

A change in the nonredundant symmetry coordinates can be transformed to a change in the primitive coordinates, as shown in eq 19.

$$\Delta\mathbf{q} = \mathbf{S}\Delta\mathbf{d} \tag{19}$$

The vibrational analysis in eqs 9−15 carries directly over to symmetry coordinates, where the $3N - 6$ symmetry coordinates **d** replace the $3N - 6$ internal coordinates **q**. As the symmetry coordinates are just fixed linear combinations of the internal coordinates, the change corresponds to replacing the **B** and $\mathbf{B}^{-1}$ matrices by their symmetry transformed equivalent $\mathbf{B}'$ and $\mathbf{B}'^{-1}$ matrices.

$$\Delta\mathbf{d} = \mathbf{S}^t\Delta\mathbf{q} = \mathbf{S}^t\mathbf{B}\Delta\mathbf{x} = \mathbf{B}'\Delta\mathbf{x} \tag{20}$$

$$\Delta\mathbf{x} = \mathbf{B}^{-1}\Delta\mathbf{q} = \mathbf{B}^{-1}\mathbf{S}\Delta\mathbf{d} = \mathbf{B}'^{-1}\Delta\mathbf{d} \tag{21}$$

The corresponding $\mathbf{G_m}$ and $\mathbf{F_q}$ matrices expressed in symmetry coordinates will be denoted $\mathbf{G'_m}$ and $\mathbf{F_d}$.

There is considerable freedom in choosing the symmetry matrix, but it should be chosen such that the resulting $\mathbf{G'_m}$ matrix is sufficiently nonsingular. One convenient way of ensuring this is to define the symmetry matrix by the eigenvectors of the **G** matrix in eq 22.

$$\mathbf{G} = \mathbf{B}\mathbf{B}^t \tag{22}$$

The **G** matrix in eq 22 is equivalent to the $\mathbf{G_m}$ matrix in eq 10 without the mass weighting and also enters the expression for the $\mathbf{B}^{-1}$ matrix, eq 8. If the **b** vectors used for constructing **G** in eq 22 are taken to be normalized, this is equivalent to performing a principal component analysis (PCA) of the **b** vectors.[15,38] By selecting the $3N - 6$ eigenvectors corresponding to the largest eigenvalues, the determinant of the $\mathbf{G'_m}$ matrix will have its maximum value, i.e., selecting the least redundant set of coordinates. This is the same procedure used for defining delocalized internal coordinated for use in geometry optimizations.[23] Alternatively, the symmetry matrix can be determined using a rule-based combination of primitive coordinates, of which the procedure for defining natural internal coordinates (NIC) is the most well-known.[22,28] The main limitation of NIC is the difficulty in defining nonredundant coordinates for systems containing fused and/or large rings. Furthermore, it may be difficult to extend the principles for constructing NIC to general types of internal coordinates, such as for example long-range coordinates discussed below.

If the PCA extraction of symmetry coordinates is performed for all the primitive coordinates together, each symmetry coordinate will in general contain contributions from all possible types, i.e., a mixture of S, B, and T type coordinates. von Arnim and Ahlrichs have in connection with geometry optimization suggested that the "hard" and "soft" degrees of freedom should be separated.[28] This may be done by partitioning the primitive coordinates into blocks and performing a PCA on each block separately. After extracting the nonredundant combinations from the first block, the primitive coordinates in the second block are orthogonalized against the symmetry coordinates from the first block, before being subjected to a PCA, and so on for all subsequent blocks. In this procedure, there are three choices: how the primitive coordinates are partitioned into blocks, the order by which the blocks are subjected to the PCA, and the threshold on eigenvalues for deciding whether a given PCA component is used for defining a symmetry coordinate. It is advantageous if the block partitioning of coordinates is done in such a way that the PCA for each block results in a clear separation of the eigenvalues of the redundant and nonredundant combinations. If the eigenvalues form a near continuous set, the number of symmetry coordinates extracted from each block will depend on a somewhat arbitrary cutoff parameter, and the number of coordinates selected in a given block will influence the number of coordinates extracted from all subsequent blocks due to the orthogonalization. By partitioning the primitive coordinates into blocks, the $\mathbf{G'_m}$ matrix will always become more singular (smaller determinant), as a block-sequential PCA for extracting symmetry coordinates corresponds to neglecting coupling elements between some of the **b** vectors, but for many

Table 1. Vibrational Analysis for the C_{60} Molecule (60 atoms, 174 vibrational modes)

	Symmetry coordinates				Normal modes		
PCA order	N_S	N_B	N_T	$-\log(\lvert\mathbf{G'}_m\rvert)$	N_S	N_B	N_T
SBT	56.9	49.4	67.7	115	42.7	42.6	88.7
S, B, T	90	84	0	192	78.3	95.7	0
S, T, B	90	0	84	144	73.5	0	100.5
T, B, S	29	0	145	202	35.2	0	138.8

N_S, N_B, and N_T indicates the sum of squared coefficients for stretch, bend, and torsion coordinates in the symmetry coordinates and in the normal modes. $\log(\lvert\mathbf{G'_m}\rvert)$ is the logarithm of the determinant of the $\mathbf{G'_m}$ matrix; see the text for discussion.

reasonable choices it will still be well within limits that allow safe numerical handling.

All results in the present paper have been obtained with the OPLS force field,[39] with force constants calculated by the Tinker program,[40] and vibrational analysis performed with a locally modified version of the Gamess-US program.[41]

Results

The key concept in the present paper is to investigate how to define primitive coordinates and extract symmetry coordinates to provide a useful set of coordinates for describing molecular vibrations without compromising numerical accuracy. The freedom in defining symmetry coordinates is illustrated in Table 1 for the C_{60} fullerene. A total of 990 primitive coordinates (90 S, 540 B, and 360 T) can be defined for this system, while only 174 nonredundant combinations are possible. If a PCA analysis is performed on all 990 primitive coordinates together, the 174 symmetry coordinates in total contain 56.9 S, 49.4 B, and 67.7 T primitive coordinates, as defined by the sum of squares of the coefficients in the rows of the symmetry matrix in eq 18 (coefficients from the PCA eigenvectors of the $\mathbf{G}$ matrix in eq 22). If the three types of internal coordinates instead are analyzed separately in the order S, B, and T, all 90 primitive stretch coordinates are nonredundant. Once these are defined as symmetry coordinates, there are 84 nonredundant combinations of the bending coordinates, and all torsional coordinates consequently become redundant. If the PCA analysis instead is performed in the order S, T, and B, 84 nonredundant combinations of the torsional coordinates are extracted, and all bending coordinates becomes redundant. The last row in Table 1 shows that if the PCA analysis is done in the order T, B, and S, the nonredundant vibrational space is described by 29 stretching and 145 torsional coordinates. The $\log(\lvert\mathbf{G'_m}\rvert)$ values show that all four of these sets of symmetry coordinates can be handled without numerical problems. The four sets of symmetry coordinates in Table 1 represent limiting cases where all PCA vectors corresponding to numerically nonzero eigenvalues are selected for each block, which for this system corresponds to eigenvalues larger than 0.01, but other combinations of S, B, and T coordinates could also be generated.

The vibrational analysis corresponds to determining the eigenvalues and -vectors of the $\mathbf{G'_m F_d}$ matrix expressed in the selected symmetry coordinates, and the normal modes can subsequently be back-transformed to the primitive $\mathbf{q}$ coordinates by eq 19. It should be stressed that all four sets of coordinates in Table 1 give the same vibrational frequencies, but the description of the corresponding normal modes in terms of primitive coordinates is substantially different. Since the normal modes expressed in the $\mathbf{q}$ coordinates are not orthogonal, the contribution of the different types of coordinates is not necessarily the same as for the symmetry coordinates, and the composition in terms of S, B, and T coordinates is also shown in Table 1. While the nonorthogonality introduces some changes in the importance of each type of coordinate, the results show that there is considerable freedom in selecting coordinates for the vibrational analysis.

One of our main motivations for the present work was to perform vibrational analysis of proteins, where the conformational degrees of freedom are determined primarily by torsional coordinates. In order to simplify the description of the vibrations related to conformational transitions, it is of interest to perform the PCA in the order S, B, and T in order to minimize the number of torsional coordinates. We furthermore propose a separation of the torsional coordinates into four groups, defined as follows:

- Hard (HD) torsions: The two central atoms either form a (partial) double bond or are part of a (small) ring system (only five- and six-membered rings are relevant for proteins). Rotation around amide bonds belongs to this class. For three-coordinated atoms with a near-planar geometry, these torsional coordinates describe the out-of-plane movement of the central atom.
- Back-Bone (BB) torsions: The two central atoms are part of the peptide backbone but are not connected by an amide bond; i.e., these are the Ramachandran ϕ and ψ angles.
- Side-Chain (SC) torsions: The two central atoms belong to an amino acid side chain, or one of them is the backbone C_α atom, but the torsion is not of the soft type.
- Soft (SF) torsions: Rotation around single bonds connected to terminal groups in a side-chain. This class includes all torsional angles describing the rotation of $-$OH, $-$SH, $-CH_3$, and $-NH_3$ groups. These four torsional blocks are analyzed in the order HD, BB, SC, and SF when the symmetry coordinates are extracted by PCA. Table 2 shows the vibrational analysis for the angiotensin octapeptide (amino acid sequence DRVYIHPF) in an extended conformation, where 158 S, 281 B, and 411 T primitive coordinates can be generated on the basis of atomic distance criteria. The torsional coordinates can, according to the above classification, be separated into 151 T_{HD}, 90 T_{BB}, 120 T_{SC}, and 50 T_{SF} coordinates. The combined PCA treating all SBT coordinates together provides symmetry coordinates corresponding to 154.7 S, 155.3 B, and 149.0 T primitive coordinates. Performing the PCA on the block-separated coordinates gives the results in the second row in Table 2, where the number of torsional coordinates now is reduced to 56. The $\log(\lvert\mathbf{G'_m}\rvert)$ value, however, shows that the symmetry coordinates between blocks have become somewhat linearly dependent. The main reason for this is a strong coupling between the B and T_{HD} coordinates, as shown by the results in the third row of Table 2, where the T_{HD}'s are allowed to mix with the bending coordinates during the PCA. The necessity of allowing the B and T_{HD} coordinates to mix

Table 2. Vibrational Analysis for the Angiotensin Octapeptide (155 atoms, 459 vibrational modes)[a]

	Symmetry Coordinates							Normal Modes					
PCA order	N_S	N_B	$N_{T\text{-}HD}$	$N_{T\text{-}BB}$	$N_{T\text{-}SC}$	$N_{T\text{-}SF}$	$-\log(\lvert\mathbf{G'_m}\rvert)$	N_S	N_B	$N_{T\text{-}HD}$	$N_{T\text{-}BB}$	$N_{T\text{-}SC}$	$N_{T\text{-}SF}$
SBT	154.7	155.3	56.0	33.3	41.4	18.6	308	99.5	104.7	77.1	50.0	80.1	47.7
S, B, T	158	245	19	15	15	7	541	114.1	163.1	37.0	40.0	65.5	39.1
S, BT_{HD}, T	158	195.7	68.3	15	15	7	354	107.4	134.9	96.0	32.3	53.8	34.6
NIC	158	205	59	15	15	7	376	111.0	148.0	67.4	36.0	59.2	37.5

[a] N_S, N_B, $N_{T\text{-}HD}$, $N_{T\text{-}BB}$, $N_{T\text{-}SC}$, and $N_{T\text{-}SF}$ indicate the sum of squared coefficients for stretch, bend, and torsion (hard, back-bone, side-chain, soft) coordinates in the symmetry coordinates and in the normal modes. $\log(\lvert\mathbf{G'_m}\rvert)$ is the logarithm of the determinant of the $\mathbf{G'_m}$ matrix; see the text for discussion. NIC indicates natural internal coordinates, using 170° as a criterion for using improper torsional coordinates for near-planar tri-coordinated atoms.

is that the HD torsional coordinates describe the out-of-plane motion for near-planar atoms, which are strongly coupled to the corresponding in-plane bending motion. The last row in Table 2 shows the results obtained using natural internal coordinates for the analysis, where an improper torsional coordinate has been used to describe the out-of-plane motion for tricoordinated atoms when the torsional angle was larger than 170°. It is seen that the PCA-based definition of the symmetry matrix gives very similar results to using rule-based NIC but avoids the problems of defining NIC for large rings[42] and using a somewhat arbitrary parameter for deciding when to use out-of-plane coordinates instead of regular bending coordinates.

The normal-mode analysis in terms of contribution from the primitive coordinates is shown in the right-hand part of Table 2. The results in the third row show that the 15 T_{SC} coordinates make contributions corresponding to a total of 53.8 normal modes, and the seven T_{SF} coordinates make contributions to 34.6 modes. The difference between the contribution of the primitive coordinates to the symmetry coordinates and the normal modes is a measure of the nonorthogonality of the normal modes in the primitive coordinates. The significance is that the PCA can extract internal degrees of freedom that span a proportionally larger part of the vibrational space than they span in terms of symmetry coordinates.

The $\log(\lvert\mathbf{G'}_m\rvert)$ value can be used as a criterion for selecting one set of symmetry coordinates over another,[37] but other criteria are also possible. In the normal coordinate system, the $\mathbf{G'_m}^{-1}$, $\mathbf{F_d}$, and $\mathbf{G'_m F_d}$ matrices are diagonal, and symmetry coordinates that make these matrices more diagonal dominant may thus indicate that they are more natural coordinates for describing the vibrational problem. As the diagonal elements of the $\mathbf{G'_m F_d}$ matrix in the normal coordinate system are proportional to the square of the frequencies, we suggest that the difference between $\sqrt{(\mathbf{G'_m F_d})_{ii}}$ and ν_i over all or a selected range of frequencies can be used as an indicator of whether a given set of symmetry coordinates is better than other alternatives.

Table 3 shows the average diagonal and off-diagonal elements of the $\mathbf{F_d}$ and $\mathbf{G'_m F_d}$ matrices, as well as the $\sqrt{(\mathbf{G'_m F_d})_{ii}} - \nu_i$ difference averaged over all frequencies. A complete separation of the S, B, and T primitive coordinates (row 2 in Table 3) leads to a small value of the $\mathbf{G'_m}$ determinant and results in an increase in the $\mathbf{F_d}$ matrix elements by roughly 3 orders of magnitude, which clearly is unfavorable. The coordinate separation where the HD torsions are combined with the bending coordinates during the PCA, however, makes both the $\mathbf{F_d}$ and $\mathbf{G'_m F_d}$ matrices more diagonal dominant, as well as reducing the average $\sqrt{(\mathbf{G'_m F_d})_{ii}} - \nu_i$ difference by roughly a factor of 2. A separation of S, B, and T coordinates, where the HD torsions are combined with the bending coordinates for defining the symmetry matrix, thus appears to be preferable over a combined approach where all primitive coordinates are allowed to mix. The use of natural internal coordinates makes the $\mathbf{F_d}$ and $\mathbf{G'_m F_d}$ matrices even more diagonal and reduces the average $\sqrt{(\mathbf{G'_m F_d})_{ii}} - \nu_i$ difference by another a factor of 2.

Table 3. Vibrational Analysis for the Angiotensin Octapeptide (155 atoms, 459 vibrational modes)[a]

PCA order	$-\log(\lvert\mathbf{G'_m}\rvert)$	$\langle\lvert F_{ii}\rvert\rangle$	$\langle\lvert F_{ij}\rvert\rangle$	$\langle\lvert GF_{ii}\rvert\rangle$	$\langle\lvert GF_{ij}\rvert\rangle$	$\langle\lvert\sqrt{(\mathbf{G'_m F_d})_{ii}} - \nu_i\rvert\rangle$
SBT	308	0.21	0.005	0.089	0.0027	471
S, B, T	541	1227.86	14.75	0.090	0.0094	220
S, BT_{HD}, T	354	0.25	0.004	0.089	0.0013	223
NIC	376	0.26	0.002	0.089	0.0003	126

[a] NIC indicates natural internal coordinates, using 170° as a criterion for using improper torsional coordinates for near-planar tri-coordinated atoms. $\langle\lvert F_{ii}\rvert\rangle$, $\langle\lvert F_{ij}\rvert\rangle$, $\langle\lvert GF_{ii}\rvert\rangle$, and $\langle\lvert GF_{ij}\rvert\rangle$ are average diagonal and off-diagonal elements of the $\mathbf{F_d}$ and $\mathbf{G'_m F_d}$ matrices, respectively (in atomic units). The last column is the average difference between the square root of the diagonal elements of the $\mathbf{G'_m F_d}$ matrix and the frequencies (in cm^{-1}).

In large proteins, the low-frequency modes often describe large-scale movements corresponding to elongation, bending, and twisting motions of the whole system or of large domains relative to each other. In either Cartesian or internal coordinates of the type discussed above, these normal modes are described as a linear combination of many coordinates, each of which enters the normal mode with a small coefficient. The existence of such large-scale movements suggests that the primitive set of coordinates should be augmented with long-range coordinates to allow a more compact description of the low-frequency modes. One possibility is to define long-range stretch, bend, and torsion coordinates between C_α atoms of each residue, and such coordinates can be generated automatically, analogous to the corresponding short-range coordinates. We have considered a scheme based on a list of C_α atoms ordered according to the protein backbone connectivity, as illustrated in Figure 1, where long-range coordinates are defined between C_α atoms that are separated by a fixed number of residues, denoted stride. A stride of 1 means that long-range coordinates are defined between all C_α atoms in neighboring residues (1−2, 2−3, 3−4, etc.); a stride of 2 means that long-range coordinates are defined between C_α atoms in every

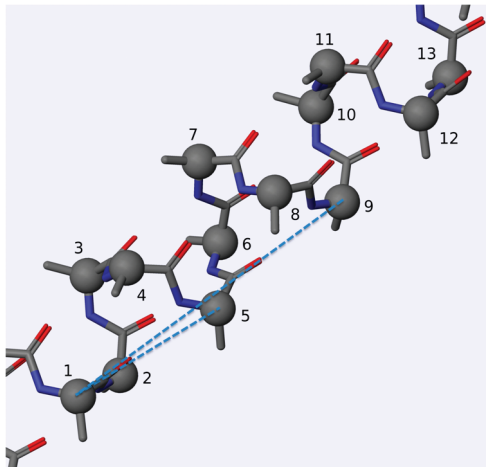

Figure 1. Illustration of a helical conformation of a polyalanine peptide with C_α atoms highlighted and hydrogen atoms omitted. Long-range stretching coordinates with a stride of 4 and 8 are indicated by dashed lines.

Table 4. The Effect of Adding Long-Range Coordinates on Vibrational Analyses for Helical and Extended Conformations of a 50 Residue Polyalanine Peptide[a]

	Helix			Extended		
stride	S_{LR}	B_{LR}	T_{LR}	S_{LR}	B_{LR}	T_{LR}
none		34.6			36.2	
1	32.8	25.9	29.6	48.8	16.7	5.1
2	5.6	12.5	9.8	0.7	17.3	4.2
3	4.6	26.1	19.4	5.0	17.5	4.8
4	1.0	14.2	19.5	0.3	15.7	3.8
5	2.4	13.4	16.4	4.5	14.5	4.8
6	1.1	15.4	12.8	0.7	14.7	4.0
7	1.6	14.9	*b*	5.1	10.0	*b*
8	0.6	*b*	*b*	1.3	*b*	*b*
9	1.2	*b*	*b*	4.4	*b*	*b*

[a] Table entries show $\langle|\sqrt{(\mathbf{G'_m F_d})_{ii}} - \nu_i|\rangle$ for the lowest five vibrational frequencies (in cm^{-1}). S_{LR}, B_{LR}, and T_{LR} indicate long-range stretch, bend, and torsional coordinates, respectively. [b] Less than five long-range coordinates added.

second residue (1−3, 3−5, 5−7, etc.); a stride of 3 means that long-range coordinates are defined between C_α atoms in every third residue (1−4, 4−7, 7−11, etc.); etc. The usefulness of stride restrictions is perhaps best illustrated by considering the α helix in Figure 1, where a stride of 4 corresponds to linking C_α atoms between residues having made approximately a full helix turn. Long-range coordinates with strides of 4, 8, etc. may thus be naturally well-suited for describing deformations of α-helix moieties. In the PCA extraction of symmetry coordinates, the long-range coordinates of a given type are collected in a separate block and are always treated before the corresponding short-range coordinates; i.e., the ordering is S_{LR}, S, B_{LR}, $B+T_{HD}$, T_{LR}, T_{BB}, T_{SC}, and T_{SF}. The PCA ensures that the nonredundant long-range coordinates are selected first, and linear combinations of short-range coordinates that span the same space are removed.

Table 4 shows the average $\sqrt{(\mathbf{G'_m F_d})_{ii}} - \nu_i$ difference for the five lowest modes for helical and extended conformations of a 50 residue polyalanine peptide as a function of adding long-range coordinates with increasing strides. For the helical conformation, the five lowest vibrational frequencies are 2.16, 2.16, 5.63, 5.67, and 6.25 cm^{-1}, while the corresponding values for the extended conformation are 0.14, 0.15, 0.34, 0.47, and 0.81 cm^{-1}, and these describe global deformation modes. In the absence of long-range coordinates, the average $\sqrt{(\mathbf{G'_m F_d})_{ii}} - \nu_i$ deviations are 34.6 and 36.2 cm^{-1}, respectively, while the corresponding values using natural internal coordinates are 45.1 and 62.9 cm^{-1} (not shown in Table 4). The addition of long-range stretching coordinates with strides larger than 1 markedly reduces the deviation, while long-range bending or torsional coordinates are less effective. For the extended conformation, the addition of long-range stretching coordinates with even strides has a larger effect than with odd strides. We have checked that a similar behavior is observed for a 49 residue polyalanine; i.e., the alternating effect is not related to the specific length of the polypeptide. For both conformations, a stride of 4 appears to be especially effective in reducing the deviation.

Table 5 shows the effects on the symmetry coordinates and normal modes by the addition of long-range stretch, bend, and torsion coordinates with a stride of 4 for helical and extended conformations of a 50 residue polyalanine peptide. For the symmetry coordinates, the long-range coordinates simply replace a corresponding number of backbone torsional coordinates, and the $\log(|\mathbf{G'_m}|)$ value changes only slightly. The main effect on the description of

Table 5. The Effect of Adding Long-Range Stretching Coordinates with a Stride of 4 on the Vibrational Analyses for Helical and Extended Conformations of a 50 Residue Polyalanine Peptide[a]

	symmetry coordinates					normal modes		average 5 mode % composition		
conformation	$N_{S\text{-}LR}$	$N_{B\text{-}LR}$	$N_{T\text{-}LR}$	$N_{T\text{-}BB}$	$-\log(\lvert\mathbf{G'}_m\rvert)$	$N_{X\text{-}LR}$	$N_{T\text{-}BB}$	$C_{X\text{-}LR}$	$C_{T\text{-}BB}$	$C_{T\text{-}other}$
helix	0	0	0	96	1239	0	241.8	0	63.9	33.3
	11	0	0	85	1225	24.7	220.2	54.9	26.9	17.1
	0	10	0	86	1244	1.5	231.3	8.0	55.4	33.8
	0	0	9	87	1226	51.7	209.0	65.4	19.9	13.9
extended	0	0	0	96	1235	0	270.7	0	98.5	1.3
	11	0	0	85	1246	19.7	244.9	17.5	80.3	1.9
	0	10	0	86	1250	4.6	252.4	33.6	60.0	5.5
	0	0	9	87	1207	254.9	148.2	100.0	0.0	0.0

[a] $N_{S\text{-}LR}$, $N_{B\text{-}LR}$, $N_{T\text{-}LR}$, and $N_{T\text{-}BB}$ indicate the number of long-range stretch, bend, torsion, and back-bone torsional coordinates, respectively, and $N_{X\text{-}LR}$ indicates the number of the particular type of long-range coordinate. $\log(|\mathbf{G'_m}|)$ is the logarithm of the determinant of the $\mathbf{G'_m}$ matrix; see the text for discussion. $C_{X\text{-}LR}$, $C_{T\text{-}BB}$, and $C_{T\text{-}other}$ indicate the average composition of the five lowest normal modes in terms of long-range, back-bone torsion, and torsion coordinates other than back-bone, respectively. The total number of normal modes is 1470. The major change by the addition of long-range coordinates is in the number of backbone torsional coordinates and torsional normal modes, and only those are listed.

the normal modes by adding long-range coordinates is also a replacement of the backbone torsional coordinates, but there is a significant difference between the type of long-range coordinate and the helical/extended conformation.

For the helical conformation, long-range stretching and torsional coordinates are more important than long-range bending coordinates. The 11 long-range stretching coordinates contribute to a total of 24.7 normal modes, as measured by the sum of squared coefficients in eq 16. The 10 long-range bending coordinates, on the other hand, only contribute to 1.5 normal modes, while the nine long-range torsional coordinates contribute to 51.7 normal modes. The long-range coordinates are primarily involved in the description of the low frequency normal modes, as illustrated by the average composition of the five lowest normal modes shown in the last three columns in Table 5. In the absence of long-range coordinates, these global deformation modes are described primarily as a combination of backbone torsional coordinates, where the weight of any primitive backbone torsional coordinate to a given normal mode is less than 1%. The addition of only 11 long-range stretching coordinates to the set of 2589 short-range primitive coordinates results in these five modes being described as a combination of long-range stretching and short-range torsional coordinates. The 11 long-range stretching coordinates describe more than 50% of the five lowest normal modes on average, and each individual long-range stretching coordinate accounts for up to ∼25% of a given normal mode. A similar effect is observed for the long-range torsional coordinates.

The extended conformation has a large number of low frequency bending and twisting modes (145 modes below 100 cm^{-1}), of which many resemble standing waves for a string. Long-range torsional coordinates are very efficient at describing these modes, and the nine long-range torsional coordinates make contributions corresponding to a total of 254.9 normal modes! The description of the five lowest normal modes accordingly changes from being a linear combination of many backbone torsional coordinates to a linear combination of essentially only the nine long-range torsional coordinates.

Consecutive helical or sheet moieties with ∼50 residues are unusual in actual proteins, and the results in Tables 4 and 5 thus primarily serve as an illustration that long-range coordinates may be advantageous for a compact description of low-frequency large-scale motions in proteins. Instead of representing the low-frequency normal modes as a linear combination of a large number of short-range primitive coordinates, these modes can instead be represented as a combination of a few long-range coordinates with large weights for describing the overall motion and number of short-range coordinates with small weights for describing the local displacements. For globular proteins, where residues may be spatially close without being close in terms of backbone bonding, additional long-range coordinates defined by distance or hydrogen bonding criteria may be advantageous. Which long-range or interstrand coordinates are most advantageous will most likely depend on the nature of the given protein. Vibrational analyses for a number of real proteins are required to probe these effects.

Summary

The present work shows that there is considerable freedom in choosing internal coordinates and in extracting nonredundant combinations of these for describing vibrational normal coordinates. For large molecular systems, like proteins, the addition of selected long-range coordinates can provide a compact description of especially the low-frequency normal modes. Vibrational normal modes calculated in general coordinates may be of use for analysis purposes, for example, for identifying possible large-scale transformations,[7] for biasing molecular dynamics simulations,[10] or for running dynamics simulations in selected internal coordinates.[43] The use of long-range coordinates may also be of interest in geometry optimizations of large flexible systems.

Acknowledgment. This work was supported by grants from the Danish Center for Scientific Computation, the Danish Natural Science Research Council, and the Villum Kann Rasmussen foundation.

References

(1) Bowman, J. M.; Christoffel, K.; Tobin, F. *J. Chem. Phys.* **1979**, *83*, 905. Heislbetz, S.; Rauhut, G. *J. Chem. Phys.* **2010**, *132*, 124129. Christiansen, O. *J. Chem. Phys.* **2004**, *120*, 2140. Christiansen, O. *J. Chem. Phys.* **2004**, *120*, 2149.

(2) Miloshevsky, G. V.; Jordan, P. C. *Structure* **2006**, *14*, 1241. Miloshevsky, G. V.; Jordan, P. C. *Structure* **2007**, *15*, 1654. Miloshevsky, G. V.; Hassanein, A.; Jordan, P. C. *J. Mol. Struct.* **2010**, *972*, 1.

(3) Bahar, I.; Lezon, T. R.; Bakan, A.; Shrivastava, I. H. *Chem. Rev.* **2010**, *110*, 1463.

(4) Kumar, P.; Joshi, D. C.; Akif, M.; Akhter, Y.; Hasnain, S. E.; Mande, S. C. *Biophys. J.* **2010**, *98*, 305.

(5) Ma, J. *Structure* **2005**, *13*, 373.

(6) Thomas, A.; Field, M. J.; Perahia, D. *J. Mol. Biol.* **1996**, *261*, 490.

(7) Tama, F.; Sanejouand, Y.-H. *Protein Eng.* **2001**, *14*, 1.

(8) Dobbins, S. E.; Lesk, V. I.; Sternberg, J. E. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 10390.

(9) Bahar, I.; Lezon, T. R.; Yang, L.-W.; Eyal, E. *Annu. Rev. Biophys.* **2010**, *39*, 23.

(10) Isin, B.; Schulten, K.; Tajkhorshid, E.; Bahar, I. *Biophys. J.* **2008**, *95*, 789.

(11) Ma, J. *Structure* **2005**, *13*, 373. Bahar, I.; Rader, A. *J. Curr. Opin. Struct. Biol.* **2005**, *15*, 586.

(12) Jackels, C. F.; Gu, Z.; Truhlar, D. G. *J. Chem. Phys.* **1995**, *102*, 3188.

(13) Njegic, B.; Gordon, M. S. *J. Chem. Phys.* **2008**, *129*, 164107.

(14) Brüschweiler, R. *J. Chem. Phys.* **1995**, *102*, 3396. Brooks, B. R.; Janezic, D.; Karplus, M. *J. Comput. Chem.* **1995**, *16*, 1522.

(15) During the course of this work, McIntosh published an essentially identical procedure for determining vibrational frequencies in general coordinates: McIntosh, D. F. *Theor. Chem. Acc.* **2010**, *125*, 177.

(16) Manson, S. A.; Law, M. M. *Phys. Chem. Chem. Phys.* **2006**, *8*, 2848. Venderell, O.; Gatti, F.; Lauvergnat, D.; Mayer, H.-D. *J. Chem. Phys.* **2007**, *127*, 184302. Stare, J.; Balint-Kurti, G. G. *J. Phys. Chem. A* **2003**, *107*, 7204.

(17) Kamiya, K.; Sugawara, Y.; Umeyama, H. *J. Comput. Chem.* **2003**, *24*, 826.

(18) Boatz, J. A.; Gordon, M. S. *J. Phys. Chem.* **1989**, *93*, 1819.

(19) Konkoli, Z.; Cremer, D. *Int. J. Quantum Chem.* **1998**, *67*, 1. Konkoli, Z.; Larsson, J. A.; Cremer, D. *Int. J. Quantum Chem.* **1998**, *67*, 11. Konkoli, Z.; Cremer, D. *Int. J. Quantum Chem.* **1998**, *67*, 29. Konkoli, Z.; Larsson, J. A.; Cremer, D. *Int. J. Quantum Chem.* **1998**, *67*, 41.

(20) Wilson, E. B., Jr. *J. Chem. Phys.* **1939**, *7*, 1047. Wilson, E. B., Jr. *J. Chem. Phys.* **1941**, *9*, 76.

(21) Wilson, E. B.; Decius, J. C.; Cross, P. C. *Molecular Vibrations*; McGraw-Hill Book Company, Inc.: New York, 1955.

(22) Fogarasi, G.; Zhou, X.; Taylor, P. W.; Pulay, P. *J. Am. Chem. Soc.* **1992**, *114*, 8191.

(23) Baker, J.; Kessi, A.; Delley, B. *J. Chem. Phys.* **1996**, *105*, 192.

(24) Pulay, P.; Fogarasi, G. *J. Chem. Phys.* **1992**, *96*, 2856.

(25) Bakken, V.; Helgaker, T. *J. Chem. Phys.* **2002**, *117*, 9160.

(26) Billeter, S. R.; Turner, A. J.; Thiel, W. *Phys. Chem. Chem. Phys.* **2000**, *2*, 2177.

(27) Peng, C.; Ayala, P. Y.; Schlegel, H. B.; Frisch, M. J. *J. Comput. Chem.* **1996**, *17*, 49.

(28) von Arnim, M.; Ahlrichs, R. *J. Chem. Phys.* **1999**, *111*, 9183.

(29) Levitt, M.; Sander, C.; Stern, P. S. *J. Mol. Biol.* **1985**, *181*, 423. Alexandrov, V.; Smith, D. M. A.; Rostkowska, H.; Nowak, M. J.; Adamowicz, L.; McCarthy, W. *J. Chem. Phys.* **1998**, *108*, 9685.

(30) Baker, J.; Puley, P. *J. Chem. Phys.* **1996**, *105*, 11100. Baker, J.; Puley, P. *J. Comput. Chem.* **2000**, *21*, 69. Swart, M.; Bickelhaupt, F. M.*Int. J. Quantum Chem.* **2006**, *106*, 2536. Maslen, P. E. *J. Chem. Phys.* **2005**, *122*, 014104.

(31) Stare, J. *J. Chem. Inf. Model.* **2007**, *47*, 840.

(32) Stare, J.; Balint-Kurti, G. G. *J. Phys. Chem. A* **2003**, *107*, 7204.

(33) Eckart, C. *Phys. Rev.* **1935**, *47*, 552.

(34) Miyazawa, T. *J. Chem. Phys.* **1958**, *29*, 246.

(35) The standard vibrational analysis in Cartesian coordinates corresponds to the **B** matrix being a unit matrix, and thus $\mathbf{G}_m = \mathbf{M}^{-1}$.

(36) Baker, J.; Kinghorn, D.; Pulay, P. *J. Chem. Phys.* **1999**, *110*, 4986.

(37) Németh, K.; Challacombe, M.; van Veenendaal, M. *J. Comput. Chem.* **2010**, *31*, 2078.

(38) A force constant weighting of the **B** matrix elements has been proposed by Lindh, R.; Bernhardsson, A.; Schütz, M. *Chem. Phys. Lett.* **1999**, *303*, 567. This weighting corresponds to a partial least-squares (PLS) approach for defining the symmetry matrix instead of a PCA.

(39) Kaminski, G. A.; Friesner, R. A.; Tirado-Rives, J.; Jorgensen, W. L. *J. Phys. Chem. B* **2001**, *105*, 6474.

(40) Tinker v5.1. http://dasher.wustl.edu/tinker/ (accessed 1/1/2010).

(41) Schmidt, M. W.; Baldridge, K. K.; Boatz, J. A.; Elbert, S. T.; Gordon, M. S.; Jensen, J. J.; Koseki, S.; Matsunaga, N.; Nguyen, K. A.; Su, S.; Windus, T. L.; Dupuis, M.; Montgomery, J. A. *J. Comput. Chem.* **1993**, *14*, 1347.

(42) Large rings in proteins are formed by disulfide bridges between cysteine residues.

(43) Pulay, P.; Paizs, B. *Chem. Phys. Lett.* **2002**, *353*, 400. Jaqaman, K.; Ortoleva, P. J. *J. Comput. Chem.* **2002**, *23*, 484. Miao, Y.; Ortoleva, P. J. *J. Comput. Chem.* **2008**, *30*, 423. Sherif, Z.; Ortoleva, P. *J. Stat. Phys.* **2008**, *130*, 669. Lee, S.-H.; Palmo, K.; Krimm, S. *Comput. Chem.* **2007**, *28*, 1107. Henriksson, K. O.; Pesonen, J. *J. Comput. Chem.* **2010**, *31*, 1882.

CT100463A