Large Amplitude Elastic Motions in Proteins from a Single-Parameter, Atomic Analysis

Monique M. Tirion*

Department of Membrane Research and Biophysics, Weizmann Institute of Science, Rehovot 76100, Israel

(Received 22 April 1996)

Normal mode analysis (NMA) is a leading method for studying long-time dynamics and elasticity of biomolecules. The method proceeds from complex semiempirical potentials characterizing the covalent and noncovalent interactions between atoms. It is widely accepted that such detailed potentials are essential to the success of NMA's. We show that a single-parameter potential is sufficient to reproduce the slow dynamics in good detail. Costly and inaccurate energy minimizations are eliminated, permitting direct analysis of crystal coordinates. The technique can be used for new applications, such as mapping of one crystal form to another by means of slow modes, and studying anomalous dynamics of large proteins and complexes. [S0031-9007(96)01063-0]

PACS numbers: 87.15.By, 87.15.He

Thermal equilibrium fluctuations of the x-ray crystal coordinates of proteins provide a basis for understanding the complex dynamics and elasticity of biological macromolecules [1]. Analysis of the normal modes of globular proteins shows an interesting anomaly. The density of the slow vibrational modes is proportional to their frequency, $g(\omega) \sim \omega$, rather than $g(\omega) \sim \omega^2$ as predicted by Debye's theory [2]. Yet, the atoms in globular proteins are packed as tightly as in solids. We show that a single-parameter potential reproduces the slow elastic modes of proteins obtained with vastly more complex empirical potentials. The simplicity of the potential permits greater insight and understanding of the mechanisms that underlie the slow, anomalous motions in biological macromolecules such as proteins.

To date, normal modes of globular proteins have been used to reproduce crystallographic temperature factors [3] and diffuse scatter [4]. Normal mode analyses (NMA's) shed light on shear and hinge motions necessary for catalytic reactions, and have been used with some success to map one crystal form of a protein into another [5]. Finally, NMA's yield macroscopic elastic moduli of large protein assemblies, based on their microscopic structure [6].

NMA studies of macromolecules are handicapped, however, by the complex phenomenological potentials used to model the covalent and nonbonded interactions between atom pairs. The necessary initial energy minimization requires a great deal of computer time and memory, and is virtually impossible for even moderately large proteins (with typically thousands of degrees of freedom) with a reasonable degree of accuracy. This inevitably leads to unstable modes which must be eliminated through elaborate methods, and which cast doubts on the validity of the analysis. Moreover, partly because the minimization is carried out *in vacuo*, the final configuration disagrees with the known crystallographic structure, complicating the interpretation of the results of NMA.

A typical example of a semiempirical potential used in molecular dynamics studies and NMA's has the form [7]

$$E_p = \frac{1}{2}\sum_{\text{bonds}} K_b(b - b_0)^2 + \frac{1}{2}\sum_{\text{angles}} K_\theta(\theta - \theta_0)^2 + \frac{1}{2}\sum_{\text{dihedrals}} K_\phi[1 + \cos(n\phi - \delta)] + \sum_{\text{nonbonded pairs}}\left[\frac{A}{r^{12}} - \frac{B}{r^6} + \frac{q_1 q_2}{Dr}\right]. \quad (1)$$

The first three terms describe the energy cost in the distortion of bond lengths, bond angles, and dihedral angles, and the last term represents steric repulsions, van der Waals attractions, and electrostatic interactions between nonbonded atoms. The various bonded constants, K_b, b_0, K_θ, etc., are specific for each type of covalent interaction, and the nonbonded constants, A and B, are specific for every type of interacting atom pairs. These constants are carefully determined from extensive theoretical and experimental studies (see Ref. [7]).

In this work, I show that in some cases the usual sophisticated potentials may be replaced by a far simpler pairwise Hookean potential, controlled by a single parameter. Such a formulation is sufficient to fully describe the anomalous low-frequency motion of large globular proteins, including time scales and eigenfrequencies, as well as displacements of atoms as predicted by eigenvectors. The simplified potential provides a very attractive alternative for the NMA of macromolecular assemblies. The derivation of the eigenvalue equation is simple, rapid, and accurate. Time-consuming and structure-distorting energy minimization are circumvented, preventing unphysical instabilities (negative eigenvalues), and the reduction in the number of fitting parameters yields to the new formulation a stronger predictive value.

The size of the computation depends on the number of internal coordinates considered. A molecule consisting of N atoms possesses $3N$ degrees of freedom, defined by bond lengths, bond angles, and rotations about bonds (or, alternatively, by the Cartesian coordinates of each of the constituent atoms). For long-chain molecules such as

proteins, bond lengths and angles are constrained within very narrow limits by the chemical bonding, while rotations about (single) bonds are much less restricted. Hence, for the analysis of slow modes one typically considers the bond lengths and angles as fixed, permitting only rotations about bonds. The latter are known as *dihedral angles* and serve as a convenient set of generalized degrees of freedom. The number of the dihedral angle coordinates is about $N/2$ and is determined by the specific makeup of the protein in question.

Given a potential energy function E_p and a set of n generalized coordinates, $\mathbf{q}'$, one first must find a stable local minimum $E_{p,\min}(\mathbf{q}' = \mathbf{q}'_0)$. The potential energy is then approximated by the quadratic form: $E_p = (1/2)\sum q_i F_{ij} q_j = (1/2)\mathbf{q}^\dagger \mathbf{F}_\mathbf{q}$ $(\mathbf{q} \equiv \mathbf{q}' - \mathbf{q}'_0)$, where $\mathbf{F}$ is the generalized force matrix

$$F_{i,j} = \left.\frac{\partial^2 E_p}{\partial q_i \partial q_j}\right|_{\mathbf{q}=0} .$$

Similarly, the kinetic energy is expressed as $E_k = (1/2) \times \dot{\mathbf{q}}^\dagger \mathbf{H} \dot{\mathbf{q}}$, where the elements of the "mass" matrix $\mathbf{H}$ are

$$H_{i,j} = \sum_{l=1}^{N} m_l \frac{\partial \mathbf{r}_l}{\partial q_i} \cdot \frac{\partial \mathbf{r}_l}{\partial q_j} .$$

$\mathbf{r}_l$ are the Cartesian coordinates of atom l, and the summation runs over all the atoms of the molecule. The $\{\partial \mathbf{r}/\partial q\}$ are moving derivatives which eliminate translational and rotational motion of the molecule as a whole.

Equations of motion are derived from Lagrange's equation, with the Lagrangian $\mathcal{L} = E_k - E_p$. Writing $q_j = \sum_k^N A_{jk} \alpha_k \cos(\omega_k t + \delta_k)$, one obtains the eigenvalue problem

$$\mathbf{FA} = \mathbf{\Lambda H A} , \tag{2}$$

subject to the normalization condition $\mathbf{A}^\dagger \mathbf{H A} = \mathbf{I}$. (This ensures that the eigenmodes diagonalize the system's Hamiltonian, $\mathcal{H} = E_k + E_p$.) The eigenfrequencies ω_i are given by the elements of the diagonal matrix $\mathbf{\Lambda}$, $\omega_i^2 = \Lambda_{ii}$, the eigenvectors are the columns of the matrix $\mathbf{A}$, and the amplitudes and phases, α_k and δ_k, are determined by initial conditions.

I replace the habitual detailed potentials, such as the one in Eq. (1), by the Hookean pairwise potential (between atoms a and b):

$$E(\mathbf{r}_a, \mathbf{r}_b) = \frac{C}{2}(|\mathbf{r}_{a,b}| - |\mathbf{r}^0_{a,b}|)^2 . \tag{3}$$

Here $\mathbf{r}_{a,b} \equiv \mathbf{r}_a - \mathbf{r}_b$ denotes the vector connecting atoms a and b, and the zero superscript indicates the given *initial* configuration. Thus, the usual minimization of the potential energy is eliminated.

Expanding to second order about $\mathbf{r}^0_{a,b}$ yields

$$E(\mathbf{r}_a, \mathbf{r}_b) = \frac{C}{2}\left(\frac{\mathbf{r}^0_{a,b} \cdot \Delta \mathbf{r}_{a,b}}{|\mathbf{r}^0_{a,b}|}\right)^2 , \tag{4}$$

where $\Delta \mathbf{r} \equiv \mathbf{r} - \mathbf{r}^0$. The strength of the potential C is a phenomenological constant, assumed to be the same for *all* interacting pairs.

The potential energy within a molecule is then given by

$$E_p = \sum_{(a,b)} E(\mathbf{r}_a, \mathbf{r}_b) . \tag{5}$$

The sum is restricted to atom pairs separated by less than $R_{\mathrm{vdW}}(a) + R_{\mathrm{vdW}}(b) + R_c$, where R_{vdW} refers to the van der Waals radii, and R_c is an arbitrary cutoff parameter which models the decay of interactions with distance. R_c determines the total number of interacting atom pairs contributing to the potential energy of the system, and is inversely related to the "bond strength" C. We shall argue below that best results are obtained with small cutoff distances. Fortunately, this tends to reduce the size of the computation.

The proposed potential seems too simplistic. For example, it is unclear how interactions between nonbonded atoms could model the 3-state "knob" potential for rotations about the backbone. The answer lies in the fact that slow vibrational modes involve coherent motion of *large* groups of atoms. The effective force opposing large scale oscillations stems from the combined effect of numerous interacting atom pairs. The sum of these interactions approaches a universal form, governed by the central limit theorem, regardless of the details of individual pair-potentials. Hence, for slow vibrations these details could be neglected.

To test this hypothesis, I compare the eigenfrequency and eigenvector data obtained using the potential of Eqs. (4) and (5) and the detailed potential of Levitt, L79 [7]. In all cases, I refer only to the slowest frequency modes. The higher frequency modes, pertaining to rapid oscillations of sidechains and small groups of atoms, require an accurate analysis at the microscopic level and could not be modeled by simplified potentials.

I performed extensive tests on the muscle protein, G-actin. This system has a molecular weight of 44 kD and contains 375 residues (3500 atoms) in a single polypeptide chain [8]. The polypeptide chain is folded so as to form two large domains joined by a narrow neck region. These two domains are partly held together by salt bridges and hydrogen bonds provided by a nucleotide (ADP) and a cation (Ca^{++}) bound in the cleft between the two domains. Complete eigenfrequency and eigenvector data exist for this system using the L79 potential [9].

Focusing first on the density of modes, let us examine the cumulative density of modes up to frequency ω, $G(\omega) = n^{-1} \int_0^\omega d\omega' \, g(\omega')$. In globular proteins $G(\omega)$ falls under a universal curve [2]. For small frequencies, $G(\omega) \sim \omega^2$. The difference from regular crystals, where $G(\omega) \sim \omega^3$, reflects the anomalous dynamics of slow vibrations in proteins.

Figure 1 shows $G(\omega)$ against ω for G-actin:ADP:Ca^{++} for the slowest 10% of the modes (138 modes). The dashed curve refers to data obtained using the standard L79 potential. Superposed are curves obtained with $R_c = 1.1$, 1.5, 2.0, and 2.5 Å, resulting in 19248, 27310,

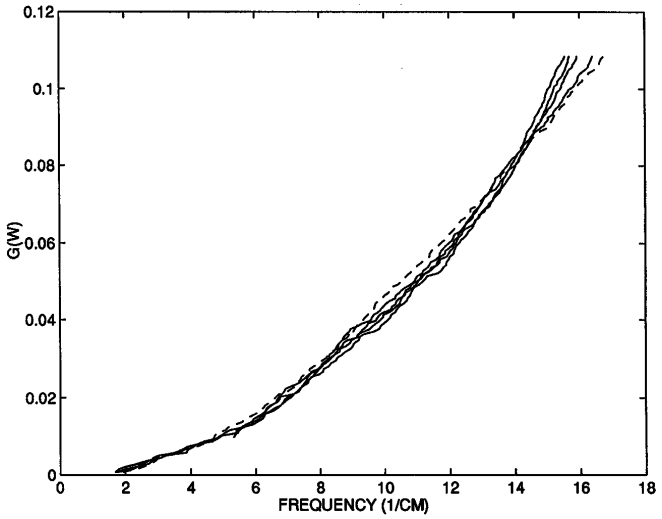

FIG. 1. The fraction of the total number of modes up to frequency ω (cm^{-1}) for the slowest 150 modes of the G-actin:ADP:Ca^{++} system. The dashed line pertains to data obtained using the L79 potential, while the four solid curves are obtained using R_C values of 1.1, 1.5, 2.0, and 2.5 Å. The $R_c = 1.1$ Å curve is nearest the dashed line at higher frequencies, with the fit progressively worsening for the higher cutoff values.

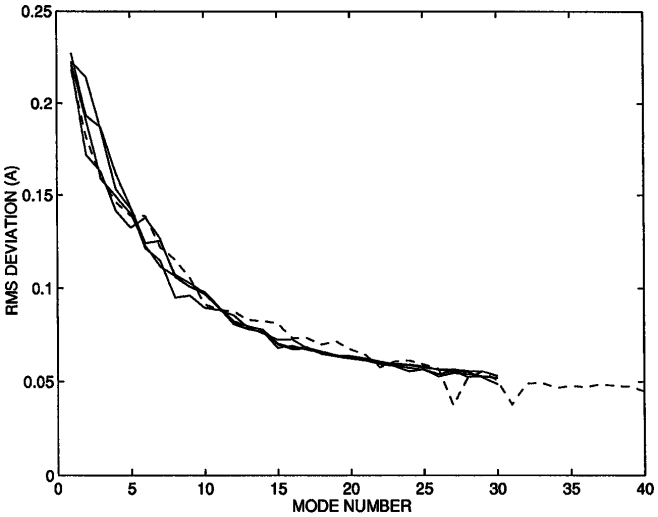

FIG. 2. The rms deviation of all mainchain C_α atoms per mode, for the slowest 30 modes. The dashed line refers to data obtained using the L79 potential, and the four solid curves are obtained with the same cutoff values, R_c, as in Fig. 1. The rms fluctuations due to the first four modes contribute over 50% to the total rms deviations of all C_α due to all modes [8]. There is no obvious improvement in choosing one particular cutoff parameter according to these data.

38654, and 51020 nonbonded interactions, respectively. To obtain optimal fits to the standard (dashed) curve, the values of C need to be adjusted to 2.49, 1.29, 0.73, and 0.47 kJ/Å^2 mol, respectively. Curiously, C seems to scale as $1/R_c^2$ rather than $1/R_c^3$. This may reflect the unusual spectral dimension, i.e., the effective dimensionality of interatomic interactions, of globular proteins, $d_s = 2$ [2]. The product CR_c^2 results in a universal "bond-strength" constant of about 3.0 kJ/mol.

Figure 1 also shows that larger cutoff values increase the curvature of the curve, and hence smaller R_c values better model the previous eigenvalue data. This emphasizes that nonbonded interactions are dominated by nearest-neighbor interactions, presumably due to screening effects.

In addition to the frequency of the modes, one can compute the root mean square (rms) displacements of atoms from equilibrium at room temperature from the eigenvector data. Figure 2 shows the rms fluctuations of all the C_α atoms as a function of the slowest 30 modes. The dashed curve shows the data obtained using the standard, L79 potential. The rms deviations decrease rapidly with mode number, indicating that the correlation length of the motion decreases as well. Superposed on the dashed curve are the four curves obtained with the same R_c and C values used in Fig. 1. Values of C predicted from the eigenfrequency data, Fig. 1, fit the eigenvector data equally well. These data also indicate no clear advantage to using larger values of the cutoff parameter, R_c. For the slowest modes, therefore, there is overall consistency and a very good match of the current results with those obtained previously, in terms of both dispersion spectra as well as rms deviations.

Theoretical temperature factors, B, used to model x-ray crystallographic temperature factors, are obtained by computing the rms fluctuations at room temperature of each C_α due to a superposition of modes. Figure 3 shows the theoretical temperature factors for each C_α of

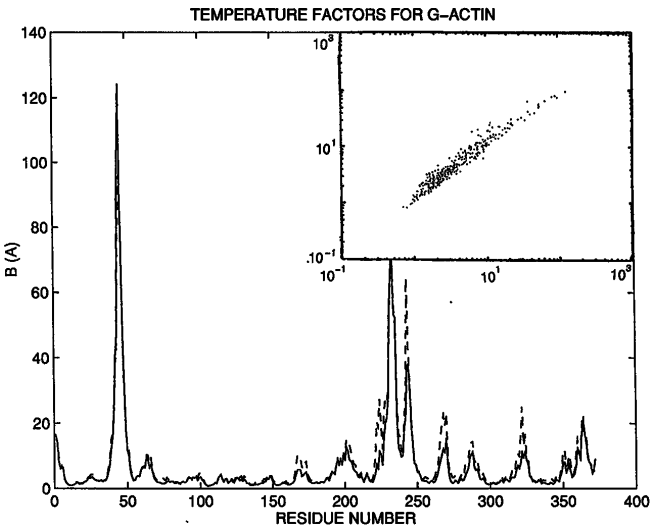

FIG. 3. Comparison of theoretical temperature factors, B, obtained with the L79 potential (dashed curve) and the potential of Eq. (1), for the G-actin:ADP:Ca^{++} system. The contributions of the 30 slowest modes are included. The inset shows the scatter plot of the two data sets: the standard potential along the ordinate, and the current simplified potential along the abscissa.

TABLE I. CPU time requirements to compute generalized eigenvalue equations.

Protein	pdb identifier	# of residues	# of coordinates	# of nonbonded interactions	CPU time (min)
Crambin	1crn	46	139	4817	0.12
Trypsin inhibitor	5pti	58	208	6529	0.45
Ribonuclease A	5rsa	124	455	14946	2.50
Lysozyme	6lyz	129	471	15834	3.10
G-actin	1atn	372	1382	37951	66.0
Myosin (HC)	1mys	780	3010	88653	390

G-actin:ADP:Ca^{++}, including the contribution of the 30 slowest modes. The data obtained with the L79 potential are shown by the dashed curve. The solid line is obtained using the current potential, with $R_c = 2.2$ Å. The very good fit argues against the need for additional parametrization in Eqs. (3)–(5).

Finally, I have tested the efficiency of NMA with the simple potential. Table I shows the central processing unit (CPU) times required for reading in coordinates and chemical formulas, indexing degrees of freedom and nonbonded interactions, and setting up the generalized eigenvalue Eq. (2). The entries show the CPU requirements on a Convex 220 for the following four proteins (the Brookhaven protein database label is given in parentheses): bovine pancreatic trypsin inhibitor (5pti), ribonuclease A (5rsa), G-actin bound with ADP and Ca^{++} (1atn), and myosin subfragment 1 bound with ADP (lmys, with the sidechains kindly modeled by Michael Lorenz). These proteins range in size from 58 residues (trypsin inhibitor) to 780 residues (myosin S1). For this table, I used a cutoff distance $R_c = 2.0$ Å. These CPU times represent improvements of 2 to 3 orders of magnitudes over earlier NMA. The main effect is due to the absence of minimizations and the faster computation of the force matrix $\mathbf{F}$ with the simple potential. It should be stressed that NMA of systems as large and poorly resolved as myosin S1 could not be undertaken with "standard" potentials, due to the accumulation of roundoff errors and distortions during minimizations.

This work demonstrates the surprising result that a single-parameter model can reproduce complex vibrational properties of macromolecular systems. The simple form of the potential dispenses with the need to perform initial energy minimizations, which are especially detrimental for NMA's due to the absence of solvent. Since the analysis proceeds directly from the crystal coordinates, it is now possible to quantitatively test whether two crystal forms of a protein, as in an "open" and "closed" configuration, are interconvertible using the slow modes as coordinates. Tests performed on a periplasmic maltodextrin binding protein indicate that the slowest modes do indeed closely map the open form into the closed form [Tirion, in preparation].

I thank Daniel ben-Avraham for useful discussions and ideas, Ken Holmes for use of computer facilities at the Max-Planck Institute for Medical Research in Heidelberg, and Michael Lorenz for the full coordinates of myosin S1. This material is based upon work supported by the National Science Foundation under Grant No. MCB-9316109.

*Electronic address: mmt@craft.camp.clarkson.edu

[1] M. Levitt, P. Stern, and C. Sander, J. Mol. Biol. **181**, 423 (1985); A. Kidera and N. Gō, J. Mol. Biol. **225**, 457 (1992); B. Brooks and M. Karplus, Proc. Natl. Acad. Sci. U.S.A. **80**, 6571 (1983).
[2] D. ben-Avraham, Phys. Rev. B **47**, 14 559 (1993).
[3] R. Diamond, Acta Crystallogr. **A46**, 425 (1990); A. Kidera and N. Gō, J. Mol. Biol. **225**, 457 (1992).
[4] P. Faure *et al.,* Nature Struct. Biol. **1**, 124 (1994).
[5] O. Marques and Y.-H. Sanejouand, Proteins **33**, 557 (1995); M. M. Tirion, D. ben-Avraham, M. Lorenz, and K. C. Holmes, Biophys. J. **68**, 5 (1995).
[6] D. ben-Avraham and M. M. Tirion, Biophys. J. **68**, 1231 (1995).
[7] S. J. Weiner *et al.,* J. Am. Chem. Soc. **106**, 765 (1984); B. R. Brooks *et al.,* J. Comp. Chem. **4**, 187 (1983); M. Levitt, J. Mol. Biol. **168**, 595 (1983).
[8] W. Kabsch *et al.,* Nature (London) **347**, 37 (1990).
[9] M. M. Tirion and D. ben-Avraham, J. Mol. Biol. **230**, 186 (1993).