Proc. Natl. Acad. Sci. USA
Vol. 87, pp. 3718–3722, May 1990
Biophysics

Refinement of protein dynamic structure: Normal mode refinement

(protein dynamics/x-ray crystallographic refinement of protein structure/Debye–Waller factor)

AKINORI KIDERA* AND NOBUHIRO GŌ†

*Protein Engineering Research Institute, Furuedai, Suita 565, Japan; and †Department of Chemistry, Faculty of Science, Kyoto University, Kyoto 606, Japan

Communicated by Hans Frauenfelder, February 20, 1990

ABSTRACT **An x-ray crystallographic refinement method, referred to as the normal mode refinement, is proposed. The Debye–Waller factor is expanded in terms of the effective normal modes whose amplitudes and eigenvectors are experimentally determined by the crystallographic refinement. In contrast to the conventional method, the atomic motions are treated generally as anisotropic and concerted. This method is assessed by using the simulated x-ray data given by a Monte Carlo simulation of human lysozyme. In this article, we refine the dynamic structure by fixing the average static structure to exact coordinates. It is found that the normal mode refinement, using a smaller number of variables, gives a better *R* factor and more information on the dynamics (anisotropy and collectivity in the motion).**

X-ray crystallography is recognized as one of the most powerful techniques for elucidating not only the static but also the dynamic structure of proteins (1, 2). X-ray diffraction data from a protein crystal provide information about the protein dynamics in the form of the isotropic temperature factor B_j for a nonhydrogen atom j, which is related to the mean-square fluctuation $\langle \Delta r_j^2 \rangle$ from the average coordinates by

$$B_j = (8\pi^2/3)\langle \Delta r_j^2 \rangle. \quad [1]$$

The temperature factor is determined by the least-squares refinement (3) of the x-ray structure, which minimizes a weighted sum of the residual

$$\sum_{\mathbf{h}} w(\mathbf{h})[|F_{obs}(\mathbf{h})| - |F_{cal}(\mathbf{h})|]^2, \quad [2]$$

where $\mathbf{h} = (h, k, l)$ refers to reciprocal-lattice points of the crystal and $|F_{obs}|$ and $|F_{cal}|$ denote the observed and calculated structure factor amplitudes, respectively. By assuming no correlations in the atomic motions belonging to different unit cells of the crystal, F_{cal} has the expression

$$F_{cal}(\mathbf{h}) = \sum_j f_j(\mathbf{h}) \exp(2\pi i \mathbf{h}\cdot\langle \mathbf{r}_j \rangle)\langle \exp(2\pi i \mathbf{h}\cdot\Delta\mathbf{r}_j)\rangle, \quad [3]$$

where f_j is the atomic structure factor usually given by four Gaussian functions and $\langle \mathbf{r}_j \rangle$ is the average coordinate. The dynamic part of the structure factor $\langle \exp(2\pi i \mathbf{h}\cdot\Delta\mathbf{r}_j) \rangle$ is referred to as the Debye–Waller factor and is usually approximated by the isotropic temperature factor in the form of

$$\langle \exp(2\pi i \mathbf{h}\cdot\Delta\mathbf{r}_j) \rangle \sim \exp[-B_j(|\mathbf{h}|/2)^2]. \quad [4]$$

The approximation of Eq. 4 is based on the assumption that the atomic motion is (*a*) harmonic and (*b*) isotropic. In the refinement procedure, isotropic temperature factors B_j are treated as independent variables.‡ This means that the atomic motions are (*c*) independent of each other. When describing the dynamic structure of real proteins, however, these approximations, especially that of isotropic motion, could be a serious limitation in improving the refinement (4). The number of observable experimental data $|F_{obs}|$ is not large enough for adopting the anisotropic temperature factor, which requires six times more parameters in F_{cal} than the number required in the isotropic case.

In the conventional refinement methods for determining the average static structure, a model of the atomic structure such as stereochemical knowledge (3) or conformational energy (5) has supplemented the limited diffraction data. In the same way, a model of dynamics should be used as supplemental information in the refinement of the dynamic structure. In this article, we describe the successful use of information that can be obtained about the dynamics of proteins from the normal mode analysis.

The basic idea of the normal mode analysis is to express the dynamics of a protein as a superposition of collective motions called normal modes (6–8); i.e., an arbitrary instantaneous displacement of atom j is written in terms of the normal mode variable σ_m by

$$\Delta r_{jk}(t) = \sum_m u_{jkm}\sigma_m(t) \ (k = 1, 2, \text{or } 3). \quad [5]$$

The coefficients u_{jkm} express a pattern of a collective motion of atoms in the mth normal mode. They are calculated theoretically from a Hessian matrix of the conformational energy at a minimum energy structure by using the formulation of the eigenvalue equation described below. They are orthonormalized as follows:

$$\sum_j \sum_k m_j u_{jkm} u_{jkn} = 1 \quad \text{if } m = n$$
$$= 0 \quad \text{if } m \neq n. \quad [6]$$

Here m_j is the mass of the jth atom. If the conformational energy surface can be approximated by a multidimensional parabola within the range of thermal fluctuation (the assumption of harmonicity), the motions of normal mode variables should be uncorrelated. Their variance is related to the corresponding angular frequency $2\pi\nu_m$, which is given as the eigenvalue $[(2\pi\nu_m)^2]$, and their covariance should therefore vanish.

$$\langle \sigma_m \sigma_n \rangle = k_B T/(2\pi\nu_m)^2 \quad \text{if } m = n$$
$$= 0 \quad \text{if } m \neq n, \quad [7]$$

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

‡The actual refinement procedure imposes the restraint that two *B* values for consecutive atoms connected by a covalent bond cannot be extremely different from each other. A part of cooperativity in the motion is thus recovered by this restraint.

where k_B is the Boltzmann constant and T is the temperature at which the x-ray data were collected. It has been shown that a relatively small number of normal modes have a dominant contribution to the atomic fluctuation. This means that the summation in Eq. 5 can be limited to a relatively small number of terms.

The real energy surface is now known to deviate significantly from a multidimensional parabola (anharmonicity) (9–13). Even in such a situation, Eq. 5 would remain valid in the sense that a significant part of atomic fluctuation can still be described in terms of a relatively small number of normal mode variables. However, because of the anharmonicity, the normal mode variables are no longer independent. In particular, the covariance of Eq. 7 may not vanish.

Here we introduce the concept of "effective normal modes," which are described in terms of the effective normal mode variables τ_m and their coefficients v_{jkm}. τ_m and v_{jkm} satisfy the same equation as Eq. 5, but σ_m and u_{jkm} have been replaced with τ_m and v_{jkm}, respectively. They are defined by requiring that covariance of the effective normal mode variables vanish even for the anharmonicity mentioned above. The coefficients are orthonormalized similarly to Eq. 6. Below we describe a formulation by which coefficients v_{jkm} and the variance are determined experimentally in the process of x-ray crystallographic refinement.

In terms of the effective normal mode variables, the Debye–Waller factor is given (instead of by Eq. 4, which is used in the conventional method) by

$$\langle \exp(2\pi i \mathbf{h} \cdot \Delta \mathbf{r}_j) \rangle \sim \exp\left[-2\pi^2 \sum_k \sum_l h_k h_l \sum_m v_{jkm} v_{jlm} \langle \tau_m^2 \rangle \right] \quad (k, l = 1, 2, \text{or } 3), \quad [8]$$

where h_k and h_l are the kth and lth components of $\mathbf{h}$, respectively. To derive this expression, we made the assumption of harmonicity of the atomic motion as is assumed in Eq. 4. However, unlike in Eq. 4, atomic motions are treated here generally as anisotropic and concerted. We call this expression the "normal mode Debye–Waller factor" and the refinement based on this expression the "normal mode refinement."

Both the effective coefficients v_{jkm} and the theoretical coefficients u_{jkm} satisfy the same orthonormalization condition of Eq. 6, so that they should be related by an orthogonal transformation:

$$v_{jkm} = \sum_n u_{jkn} \rho_{nm}, \quad [9]$$

where ρ_{nm} is an element of the orthogonal transformation matrix. Because v_{jkm} should be similar to u_{jkm}, we can use the latter as an initial guess of the former. By substituting Eq. 9 into Eq. 8, we have the following equation:

$$\langle \exp(2\pi i \mathbf{h} \cdot \Delta \mathbf{r}_j) \rangle \sim \exp\left[-2\pi^2 \sum_k \sum_l h_k h_l \sum_m \sum_n u_{jkm} u_{jln} \langle \sigma_m \sigma_n \rangle \right] \quad (k, l = 1, 2, \text{or } 3), \quad [10]$$

where

$$\langle \sigma_m \sigma_n \rangle = \sum_p \rho_{mp} \rho_{np} \langle \tau_p^2 \rangle \quad [11]$$

In the process of the normal mode refinement, the variances and covariances of the normal mode variables $\langle \sigma_m \sigma_n \rangle$ in Eq. 10 are treated as parameters to be optimized in the residual of Eq. 2. It is noted from Eq. 11 that ρ_{mp} and $\langle \tau_p^2 \rangle$ are the eigenvectors and eigenvalues of the matrix whose elements are $\langle \sigma_m \sigma_n \rangle$. Since the variances of the effective normal modes $\langle \tau_m^2 \rangle$ are greater than or equal to zero, the matrix of $\langle \sigma_m \sigma_n \rangle$ should be positive semidefinite.

It is noted that the effective normal modes depend on the temperature at which the experiment is done because they are determined by the experiment. In this context, the method of the effective normal modes can be called one of quasi-harmonic models (14, 15).

As the first stages of developing the dynamic structure refinement technique, we assess the normal mode refinement method with simulated x-ray data. The simulation is free from errors due to lattice disorder and diffusion so that one can focus on the errors in the refinement due solely to the internal atomic fluctuation. A Monte Carlo simulation of conformational dynamics of a protein, human lysozyme taken as the example, is carried out to generate simulated x-ray diffraction data. Then the normal mode refinement method is applied to the simulated data. In this article, in order to focus on the dynamic structure, we have kept the average coordinates at the exact values obtained from the Monte Carlo simulation during the refinement.

METHODS

Monte Carlo Simulation and Generation of the X-Ray Diffraction Data. Human lysozyme [1LZ1 (16) in the Brookhaven Protein Data Bank] consisting of 130 amino acid residues is chosen as the example. Monte Carlo simulation of lysozyme is carried out at 300 K in the dihedral angle space of ECEPP (17) with the scaled collective variables corresponding to the normal modes (18). No solvent molecules are included. This simulation is started from a low conformational energy structure obtained by the following two steps: (*a*) reguralization of the coordinates so that the bond lengths and bond angles take the standard values and then (*b*) minimization of the conformational energy with the Newton–Raphson method using both gradient and Hessian values of the conformational energy (19). Here the distance information from the x-ray structure is used as a restraint to avoid a large deformation by the energy minimization *in vacuo*. The root-mean-square displacement of the nonhydrogen atoms in the starting structure from the x-ray structure is 1.25 Å.

Two thousand structures are sampled at every 100 steps from the records of the 10,000th step to the 210,000th step and are used to generate the diffraction data. Corresponding to Eq. 3, $|F_{obs}|$ is given by

$$|F_{obs}(\mathbf{h})| = \left| \left\langle \sum_j f_j(\mathbf{h}) \exp(2\pi i \mathbf{h} \cdot \mathbf{r}_j) \right\rangle \right|, \quad [12]$$

where $\langle\ \rangle$ denotes the average over the 2000 sampled structures that are superposed to the x-ray structure and the sum is taken over all nonhydrogen atoms of four molecules in a unit cell generated by a symmetric operation of the space group $P2_12_12_1$. This averaging is done in the reciprocal space. The total number of 18,794 reciprocal lattice points $\mathbf{h} = (h, k, l)$ of the $P2_12_12_1$ symmetry between 1.5-Å and 10-Å resolution are considered.

Normal Mode Analysis. A normal mode analysis is carried out around the average structure of the Monte Carlo simulation by the following four steps. (*a*) The 2000 sampled structures of the Monte Carlo simulation are averaged. (*b*) The average structure is regularized as also done in the preparation of the initial structure for the simulation. This is necessary because the averaged structure does not necessarily keep the standard bond lengths and bond angles. (*c*) The conformational energy of the regularized structure is minimized to obtain a minimum energy structure whose Hessian matrix of conformational energy is positive definite.

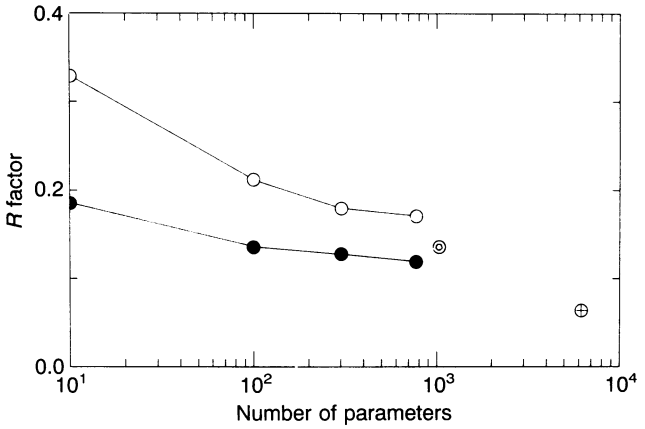

FIG. 1. R factor of Eq. **15** vs. the values of M (the number of parameters used in the model of dynamic structure). ◎, Result of the conventional isotropic B factor model. Whereas values of B_j were determined by optimizing the target functions of Eq. **2** in the real refinement procedure, we used here the exact values of $\langle r_j^2 \rangle$ obtained from the record of the simulation to calculate the B values in Eq. **4**. The use of such B values ensures that the value of the R factor obtained is the best one in the conventional refinement method. ⊕, Result of the anisotropic B factor model. As in the isotropic B factor model, the exact values of $\langle r_{jk} r_{jl} \rangle$ (k, l = 1, 2, or 3) obtained from the record of the simulation were used to calculate the anisotropic B values. ○, Initial value of the R factor, when the theoretically calculated coefficients u_{jkm} and variances were used in the normal mode Debye–Waller factor of Eq. **8**. When the values of variances were allowed to change (but the coefficients u_{jkm} were not allowed to change temporally in this article) so as to optimize the target function of Eq. **2**, the R factors were improved (●).

The root-mean-square displacement of nonhydrogen atoms in the minimum energy structure from the average structure is 0.65 Å. (*d*) The normal mode analysis in the form of the generalized eigenvalue equation in the dihedral angle space (771 dihedral angles in lysozyme) (6) is done at the minimum energy structure obtained in step *c*.

$$\mathbf{F\Omega} = \mathbf{H\Omega\Lambda} \quad \text{with} \quad \mathbf{\Omega}^t\mathbf{H\Omega} = \mathbf{I}, \qquad [13]$$

where $\mathbf{F}$ (= $\{\partial^2 E/\partial\theta_p\partial\theta_q\}$) is the Hessian matrix of the conformational energy E in terms of dihedral angles θ_p and θ_q, $\mathbf{H}$ (= $\{h_{pq}\}$; $h_{pq} = \Sigma_j m_j(\partial \mathbf{r}_j/\partial\theta_p)\cdot(\partial \mathbf{r}_j/\partial\theta_q)$) is the coefficient matrix of the kinetic energy Lagrangian in the dihedral angle space, m_j is the mass of atom j, $\mathbf{\Lambda}$ (= $\{(2\pi\nu_m)^2\}$) and $\mathbf{\Omega}$ (= $\{\omega_{pm}\}$) are the eigenvalue and eigenvector matrices, respectively, and $\mathbf{I}$ is the identity matrix. Then the eigenvectors $\{\omega_{pm}\}$ in the dihedral angle space are converted to those in the Cartesian coordinate space by

$$u_{jkm} = \sum_p (\partial r_{jk}/\partial\theta_p)\omega_{pm}. \qquad [14]$$

An analytical expression of the coefficient of this conversion has been given elsewhere (20). These values of $2\pi\nu_m$ and u_{jkm} are used as the initial guess in the refinement.

Refinement of Dynamic Structure. Eq. **2** is minimized by the quasi-Newton method (DMING1 of FACOM Scientific Subroutine Library SSL II), which uses only the gradients of Eq. **2**. In this refinement, the average coordinates are fixed to the exact values obtained from the Monte Carlo simulation, and the weighting factors $w(\mathbf{h})$ in Eq. **2** are all set to unity. The extent of summation in Eq. **5** is taken over the normal modes with the M lowest frequencies, where the values of M examined are 771, 300, 100, and 10. In each case, there are M diagonal and $M(M-1)/2$ off-diagonal independent elements of the variance and covariance matrix that are to be treated as parameters of optimization of the target function. Whereas the original idea of the normal mode refinement was to use a relatively small value of M and both the diagonal and off-diagonal elements of the variance and covariance matrix in the optimization, in this article we will restrict ourselves to the use of only the diagonal elements. This means that, as to the coefficients, we will rely on the theoretically calculated ones (i.e., $v_{jkm} = u_{jkm}$) and that we will determine only the

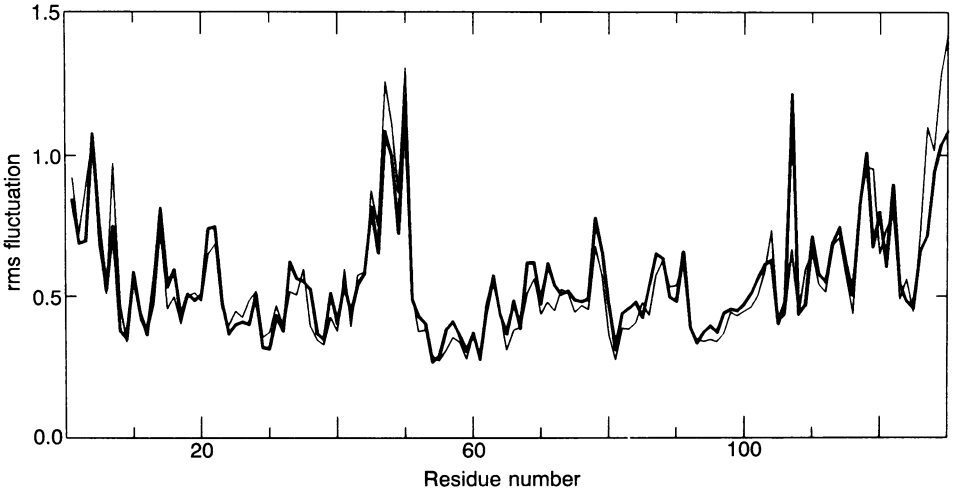

FIG. 2. Root-mean-square (rms) fluctuation $\langle r_j^2 \rangle$ averaged over all the nonhydrogen atoms within a residue. The thin curve was determined from the Monte Carlo simulation, and the thick one was calculated from the optimized values of the variance with M = 100.

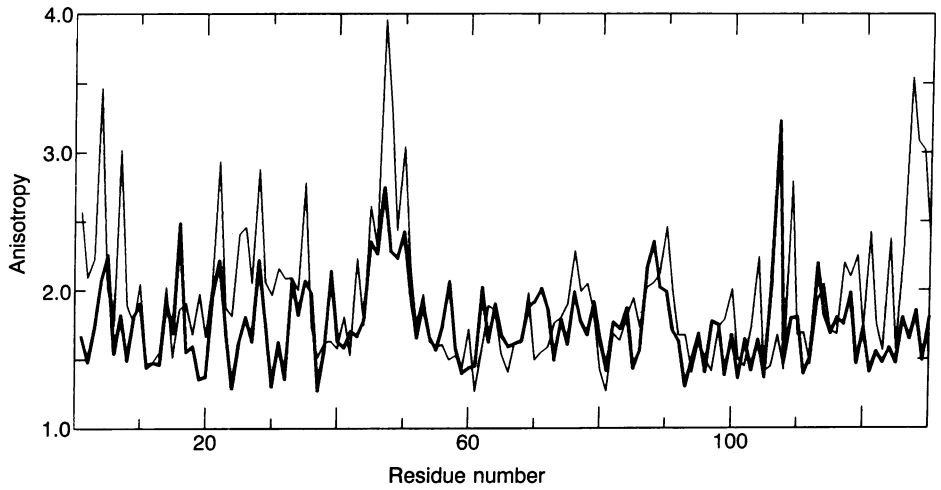

FIG. 3. Anisotropy of the distribution of the atomic motion represented by the axial ratio defined by $2\lambda_1/(\lambda_2 + \lambda_3)$ averaged within a residue. The thin curve was calculated from the Monte Carlo simulation, and the thick one was calculated from the optimized value of the variance with $M = 100$.

corresponding angular frequencies experimentally by the refinement. We will demonstrate that the normal mode refinement method is a powerful one, even in this restricted use. To clarify the characteristics of this method, the dependence of the results on different choices of the value of M, including a rather large one, is examined.

RESULTS AND DISCUSSION

The progress of the crystallographic refinement is usually assessed by the R factor:

$$R = \frac{\sum_{\mathbf{h}} ||F_{obs}(\mathbf{h})| - |F_{cal}(\mathbf{h})||}{\sum_{\mathbf{h}} |F_{obs}(\mathbf{h})|}. \quad [15]$$

Fig. 1 shows the values of the R factor for several refinement runs against M (the number of normal modes used in the refinement—i.e., the number of variables in the model of the dynamic structure). The conventional refinement method gives $R = 13.59\%$. In the conventional method, the isotropic B factors of Eq. **4** are used, and therefore the number of variables for the dynamic structure is the same as the number of nonhydrogen atoms in lysozyme (i.e., $M = 1029$). If a 6 times larger number of adjustable parameters is used for the anisotropic B factors (i.e., $M = 6174$), R can be reduced to 6.45%, which gives the lowest limit of the R factor in the harmonic approximation. The initial values of the R factor, when the normal mode Debye–Waller factor of Eq. **8** with theoretically calculated coefficients and variances is used, are already quite reasonable. For $M = 100$, it is already close to 20% without using any adjustable parameters. When the values of variances are optimized, the R factors are significantly improved. They are 11.96% for $M = 771$, 12.78% for $M = 300$, 13.58% for $M = 100$, and 18.49% for $M = 10$. These values for $M \geqq 100$ are smaller than the best value that can be attained in the conventional refinement method.

Atomic fluctuations calculated from the values of optimized variances reproduce those determined from the Monte Carlo simulation. Fig. 2 compares the root-mean-square

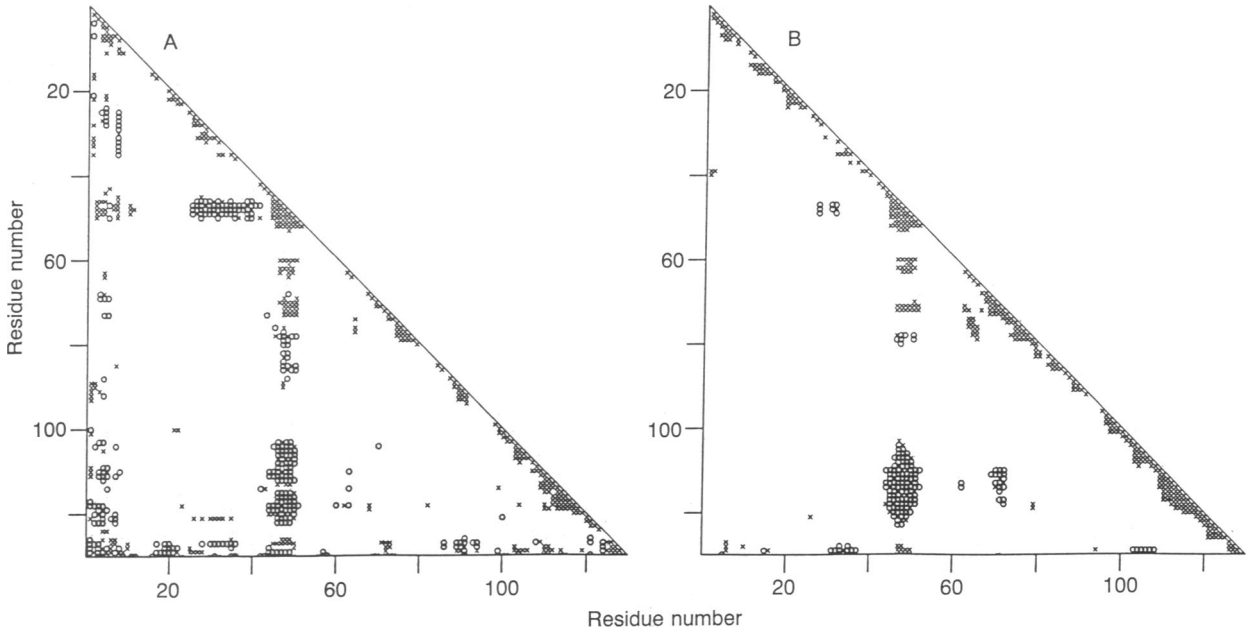

FIG. 4. Covariance in the atomic motion $\langle \Delta\mathbf{r}_i \cdot \Delta\mathbf{r}_j \rangle$ between two α-carbon atoms (Eq. **16**). (*A*) Covariances calculated from the Monte Carlo simulation. (*B*) Covariances calculated by the optimized value of variance with $M = 100$. ×, Positive covariance >0.1 Å^2 (1 Å = 0.1 nm); ○, negative covariance <−0.1 Å^2.

fluctuation given by the simulation with those of the refinement.

These improvements over the conventional refinement are due primarily to the inclusion of anisotropy in the Debye–Waller factor. The normal modes express the anisotropic motion correctly, even with a small number of parameters. Anisotropy of the ellipsoidal distribution of atomic motion shown in Fig. 3 indicates the importance of the anisotropic motion in the protein dynamics. Anisotropy in atom j is represented in the form of the axial ratio defined by $2\lambda_1/(\lambda_2 + \lambda_3)$, where λ_1, λ_2, and λ_3 are the lengths of the first, the second, and the third principal axes, respectively, of the thermal ellipsoid calculated from the Monte Carlo simulation data and also by the effective mode variances. The real distribution is far from the isotropic value of 1.0. When averaged over all nonhydrogen atoms, the ratio from the record of simulation is 1.96 and that from the optimized variance is 1.77, a good agreement.

The collectivity in the atomic motion in a protein is reflected in the following covariance:

$$C_{ij} = \langle \Delta \mathbf{r}_i \cdot \Delta \mathbf{r}_j \rangle \quad [16]$$

Fig. 4 shows the covariances between two α-carbon atoms, one from the Monte Carlo simulation and the other from the optimized value of variance. The correlation coefficient between the two figures is 0.72. The important and function-related motion in lysozyme, the hinge-bending motion, which has the negative correlation in the motion between the residues around Ala-47 and those around Val-110, is also clearly given by the refinement model.

In conclusion, the normal mode refinement, using a smaller number of variables, is shown to give better a R factor and more information on the dynamics (anisotropy and collectivity in the motion).

Finally we refer to an application of the normal mode refinement to real diffraction data. This method is intended to replace the final stage of the refinement procedure, which uses independent temperature factors. The process of the normal mode refinement will be in three steps. (*a*) Minimization of the conformational energy of an average static structure $\langle \mathbf{r}_j \rangle$, which is given as a result of the preceding step using a uniform value for all temperature factors. (*b*) Normal mode analysis at the minimum energy conformation obtained after step *a* to give the theoretical normal mode frequencies ν_m and eigenvectors u_{jkm}. The low frequency normal modes of the minimum energy structure, which are to be used in the refinement, can be regarded as being similar to those of the average static structure $\langle \mathbf{r}_j \rangle$ with a good approximation because the conformational space spanned by the low frequency normal modes is determined mainly by the large scale structure like packing topology and is insensitive to local structural differences. (*c*) Optimization of the residual of Eq. 2 in terms of the variables $\langle \mathbf{r}_j \rangle$ and $\langle \sigma_m \sigma_n \rangle$, whose initial guess is given by Eq. 7.

We thank Dr. M. Matsushima for helpful discussions. Computation was done at the Protein Engineering Research Institute and at the computer centers of Kyoto University and the Institute for Molecular Science. This work was supported by grants to N.G. from the Ministry of Education, Science and Culture, Japan, and from the Science and Technology Agency, Japan.

1. Petsko, G. A. & Ringe, D. (1984) *Annu. Rev. Biophys. Bioeng.* **13**, 331–371.
2. Frauenfelder, H. (1989) *Int. J. Quantum Chem.* **35**, 711–715.
3. Hendrickson, W. A. (1985) *Methods Enzymol.* **115**, 252–270.
4. Kuriyan, J., Petsko, G. A., Levy, R. M. & Karplus, M. (1986) *J. Mol. Biol.* **190**, 227–254.
5. Jack, A. & Levitt, M. (1978) *Acta Crystallogr. Sect. A* **34**, 931–935.
6. Gō, N., Noguti, T. & Nishikawa, T. (1983) *Proc. Natl. Acad. Sci. USA* **80**, 3696–3700.
7. Brooks, B. & Karplus, M. (1983) *Proc. Natl. Acad. Sci. USA* **80**, 6571–6575.
8. Levitt, M., Sander, C. & Stern, P. S. (1983) *Int. J. Quantum Chem. Quantum Biol. Symp.* **10**, 181–199.
9. Noguti, T. & Gō, N. (1989) *Proteins* **5**, 97–103.
10. Noguti, T. & Gō, N. (1989) *Proteins* **5**, 104–112.
11. Noguti, T. & Gō, N. (1989) *Proteins* **5**, 113–124.
12. Noguti, T. & Gō, N. (1989) *Proteins* **5**, 125–131.
13. Noguti, T. & Gō, N. (1989) *Proteins* **5**, 132–138.
14. Kobayashi, M. & Tadokoro, H. (1977) *J. Chem. Phys.* **66**, 1258–1265.
15. Levy, R. M., Srinivasan, A. R., Olson, W. K. & McCammon, J. A. (1984) *Biopolymers* **23**, 1099–1112.
16. Blake, C. C. F., Pulford, W. C. A. & Artymiuk, P. J. (1983) *J. Mol. Biol.* **167**, 693–723.
17. Némethy, G., Pottle, M. S. & Scheraga, H. A. (1983) *J. Phys. Chem.* **87**, 1883–1887.
18. Noguti, T. & Gō, N. (1985) *Biopolymers* **24**, 527–546.
19. Noguti, T. & Gō, N. (1983) *J. Phys. Soc. Jpn.* **52**, 3685–3690.
20. Noguti, T. & Gō, N. (1983) *J. Phys. Soc. Jpn.* **52**, 3283–3288.