Polymer Vol. 37 No. 10, pp. 1805–1813, 1996

0032-3861/96/$15.00+0.00

Characterizing large correlated fluctuations of macromolecular conformations in torsion-angle space using the multivariate wrapped-Gaussian distribution

Bruce W. Church and David Shalloway*
Biophysics Program, Section of Biochemistry, Molecular and Cell Biology, Cornell University, Ithaca, NY 14853, USA
(Received 30 May 1995; revised 3 July 1995)

We describe a multivariate wrapped-Gaussian distribution that can be used to model statistical distributions expressed in periodic angular variables even in the presence of significant angle–angle correlations. Using amino acid and peptide examples, we show how this distribution can be used to characterize accurately conformational fluctuations in torsion-angle space. The method permits more accurate characterization of large-scale peptide and polymer fluctuations than the standard harmonic or quasiharmonic method. This can be used to increase the efficiency of Monte Carlo sampling and simulated annealing in torsion-angle space. Copyright © 1996 Elsevier Science Ltd.

(Keywords: macromolecular conformations; correlated fluctuations; wrapped-Gaussian distribution)

INTRODUCTION

The motions of large polymers, proteins and other complex macromolecules are primarily stochastic and can be characterized by a fluctuation tensor whose eigenvectors and eigenvalues determine the principal axes and moments of fluctuation, respectively. For the small motions that occur at low temperatures, this tensor can be calculated using harmonic or quasiharmonic methods. These model the conformational probability distribution $p(\boldsymbol{R})$ as a multi-dimensional Gaussian:

$$p_{\mathrm{G}}(\boldsymbol{R}) = \det{}^{-1}(\sqrt{2\pi}\boldsymbol{\Lambda}) \exp[-(\boldsymbol{R}-\boldsymbol{R}^0)(2\boldsymbol{\Lambda}^2)^{-1}(\boldsymbol{R}-\boldsymbol{R}^0)] \tag{1}$$

Here vector $\boldsymbol{R} \equiv \{\boldsymbol{r}_i\}$ specifies the positions $\boldsymbol{r}_i$ of all the atoms in a particular conformation, $\boldsymbol{R}^0$ specifies the mean conformation of the macromolecule and $\boldsymbol{\Lambda}^2$ is the symmetric matrix specifying the components of the mean-square fluctuation. For the very small fluctuations that occur at low temperature T, $\boldsymbol{R}^0$ and $\boldsymbol{\Lambda}$ can be estimated from the first- and second-order derivatives of the Gibbs/Boltzmann distribution $\mathrm{e}^{-\beta V}$ corresponding to the molecular potential $V(\boldsymbol{R})$. This gives the *harmonic approximation*:

$$\left.\frac{\partial V(\boldsymbol{R})}{\partial \boldsymbol{R}}\right|_{\boldsymbol{R}=\boldsymbol{R}^0} = 0 \tag{2}$$

$$\left.\beta\frac{\partial^2 V(\boldsymbol{R})}{\partial \boldsymbol{R}^2}\right|_{\boldsymbol{R}=\boldsymbol{R}^0} = \boldsymbol{\Lambda}^{-2} \tag{3}$$

where $\beta \equiv (k_{\mathrm{B}}T)^{-1}$. This approximation has been used to characterize small-amplitude protein motions in terms of their eigenmodes. For example, Brooks and Karplus used it to show that the fluctuations of bovine pancreatic trypsin inhibitor are dominanted by low-frequency modes[1].

Equations (2) and (3) are accurate only when T is so low that V is approximately quadratic in the thermally sampled region. At higher temperatures, where the potential in the sampled region is anharmonic, the *quadiharmonic approximation* can be employed. It is most often used to model data sets $\{\boldsymbol{R}_\alpha : \alpha = 1, \ldots, N^d\}$ of N^d conformations $\boldsymbol{R}_\alpha$ obtained by molecular dynamics or Monte Carlo computational sampling. As in conventional statistics, we estimate $\boldsymbol{R}^0$ and $\boldsymbol{\Lambda}$ by requiring that the first and second integral moments of the data set equal the corresponding moments of the Gaussian distribution:

$$\langle \boldsymbol{R} \rangle = \int \boldsymbol{R} p(\boldsymbol{R})\, \mathrm{d}\boldsymbol{r} = \boldsymbol{R}_0 \tag{4}$$

$$\langle (\boldsymbol{R} - \langle \boldsymbol{R} \rangle)(\boldsymbol{R} - \langle \boldsymbol{R} \rangle) \rangle = \int (\boldsymbol{R} - \boldsymbol{R}^0)(\boldsymbol{R} - \boldsymbol{R}^0) p(\boldsymbol{R})\, \mathrm{d}\boldsymbol{r} = \boldsymbol{\Lambda}^2 \tag{5}$$

where angle brackets here and throughout the paper denote averages over the data set:

$$\langle f \rangle \equiv \frac{1}{N^d} \sum_{\alpha=1}^{N^d} f_\alpha \tag{6}$$

(As in statistics, for best estimation of the sample variance, the left-hand side of equation (5) should be

* To whom correspondence should be addressed

multiplied by the factor $N^d/(N^d - 1)$. However, since $N^d \gg 1$ in our application, this factor can be ignored.)

The quasiharmonic approximation in Cartesian coordinates has been widely used for polymer studies. For expample, Karplus and Kushick[2] used it to calculate the conformational entropies of polymers. Berendsen and colleagues[3] used the method to analyse the fluctuations of proteins at 300 K and to show that most (≈90%) of the fluctuation at this temperature is expressed by a small subset (≈ 5%) of the fluctuation modes.

However, these Cartesian coordinate methods are only applicable to proteins when the eigenvectors of $\mathbf{\Lambda}^2$ are tangent to the manifold of fixed bond lengths and bond angles. This will be the case for small fluctuations. However, for sufficiently large fluctuations, the motions along these eigenvectors can significantly violate bond angle and bond length constraints and do not represent physical motions. Although it is possible to project these Cartesian modes onto the manifold of physically allowed motion in the small-angle limit[4], such projections necessarily involve errors for large-angle fluctuations.

Large motions can be more compactly and accurately represented in torsion-angle space. In this case, equation (1) is replaced by:

$$p_J(\boldsymbol{\theta}) = \det{}^{-1}(\sqrt{2\pi}\mathbf{\Xi}) \exp[-(\boldsymbol{\theta} - \boldsymbol{\theta}^0)(2\mathbf{\Xi}^2)^{-1}(\boldsymbol{\theta} - \boldsymbol{\theta}^0)] \tag{7}$$

where $\boldsymbol{\theta} \equiv \{\theta_j : j = 1, \ldots, N\}$ is an angular conformation vector and $\mathbf{\Xi}^2$ is a symmetric tensor in the N-dimensional torsion-angle space. $\boldsymbol{\theta}^0$ and $\mathbf{\Xi}^2$ can be approximated using either equation (3) or equation (5) with the replacements:

$$\boldsymbol{R} \rightarrow \boldsymbol{\theta} \tag{8}$$

$$\mathbf{\Lambda} \rightarrow \mathbf{\Xi} \tag{9}$$

Although equation (7) does not respect the periodicity of the angular variables, it can be used as long as the angular fluctuations are $\ll \pi$. For example, Gō *et al.*[5] used the harmonic approximation in torsion-angle space to identify the low-temperature normal modes of bovine pancreatic trypsin inhibitor, and thereby to calculate its entropy and free energy.

An important application of these methods is Monte Carlo sampling in torsion-angle variables for computing thermodynamic properties. Efficient sampling can only be achieved if the transition function, which specifies the size scale of the random jumps, is roughly matched to the anisotropic curvature of the energy function in the sampling region. However, Monte Carlo sampling is most commonly applied using an isotropic transition function with a single size scale determined by a simple fractional move acceptance rate criterion[6]. This approach is inefficient in large-dimension polymer probably accounts for the relative unpopularity of this sampling method compared to its primary competitor, molecular dynamics[7]. Noguti and Gō[8] have analysed proteins by anisotropic Monte Carlo sampling using the harmonic approximation to determine the curvature. They showed that the use of an anisotropic transition function can improve efficiency by 50- to 500-fold. Vanderbilt and Louie[9] have described how the quasiharmonic approximation can be used to approximate an anisotropic transition function for anharmonic potentials in Cartesian space. This method has been adapted to a torsion-angle description of proteins by Shin and Jhon[10,11].

Unfortunately, even the quasiharmonic approach is not accurate at high temperatures where fluctuations are very large. The problem is that the periodicity and non-linearity of the angle variables are not properly accounted for. The high-temperature domain is particularly important for computational studies of protein folding and protein folding pathways. For example, computational analysis of the pentapeptide Met-enkephalin has shown that important pathway choices may be made at temperatures as high as 15 kcal $\text{mol}^{-1} \equiv 7500\,\text{K}$[12]. High-temperature simulations are also needed for computational refinement of macromolecular X-ray crystallography data[13]. Current algorithms for this purpose[14] typically use high-temperature molecular dynamics simulations to explore large regions of conformation space while searching for the best fit to the experimental data. Anisotropic Monte Carlo sampling in torsion-angle space might be more efficient but has not been applied because of the lack of appropriate methods for determining the angular transition function.

Given the fundamental nature of the Gaussian model (1) in (Euclidean) multivariate statistics, it is surprising that no methods to address this problem in multidimensional angular space yet exist. A number of methods have been used to describe unimodal distributions in one angular variable[15], but they do not directly generalize to the multivariate case. To address this need, we have developed a new method for calculating fluctuation tensors in torsion-angle space. Among other applications, this method can be used to characterize large high-temperature polymer and biopolymer fluctuations, for defining high-temperature 'effective' normal modes, and for calculating anisotropic Monte Carlo transition functions.

MATHEMATICAL METHODS

For illustration, we first consider a simple case: the thermal distribution of the dihedral ϕ, ψ angles of terminally blocked valine. Scatter plots obtained by Monte Carlo sampling at three temperatures are shown in *Figure 1*. (These data were generated using an adiabatic potential generated using the Moil force field[16]. Valine was blocked at the amino and carboxyl ends with methyl groups.) The Gaussian approximations to these distributions determined by the angular quasiharmonic approximation, equations (4) and (5) with (8) and (9), are represented in panels (a)–(c) by the shaded regions bounded by the contours indicating where $p_J(\boldsymbol{\theta}) = \text{e}^{-2}$. We will call these '$\text{e}^{-2}$ regions'. (These regions would contain ~ 87% of the points generated by a perfectly modelled two-dimensional Gaussian distribution.)

At low temperature (100 K) the points are tightly clustered. The correspondence of the e^{-2} region with the distribution of points in *Figure 1a* indicates that, in this case, the angular quasiharmonic approximation accurately characterizes the distribution. However, this method fails at higher temperatures, as can be seen by comparing the scatter plot resulting from molecular dynamics simulation at 300 K with the corresponding quasiharmonic e^{-2} region (*Figure 1b*). The method fails

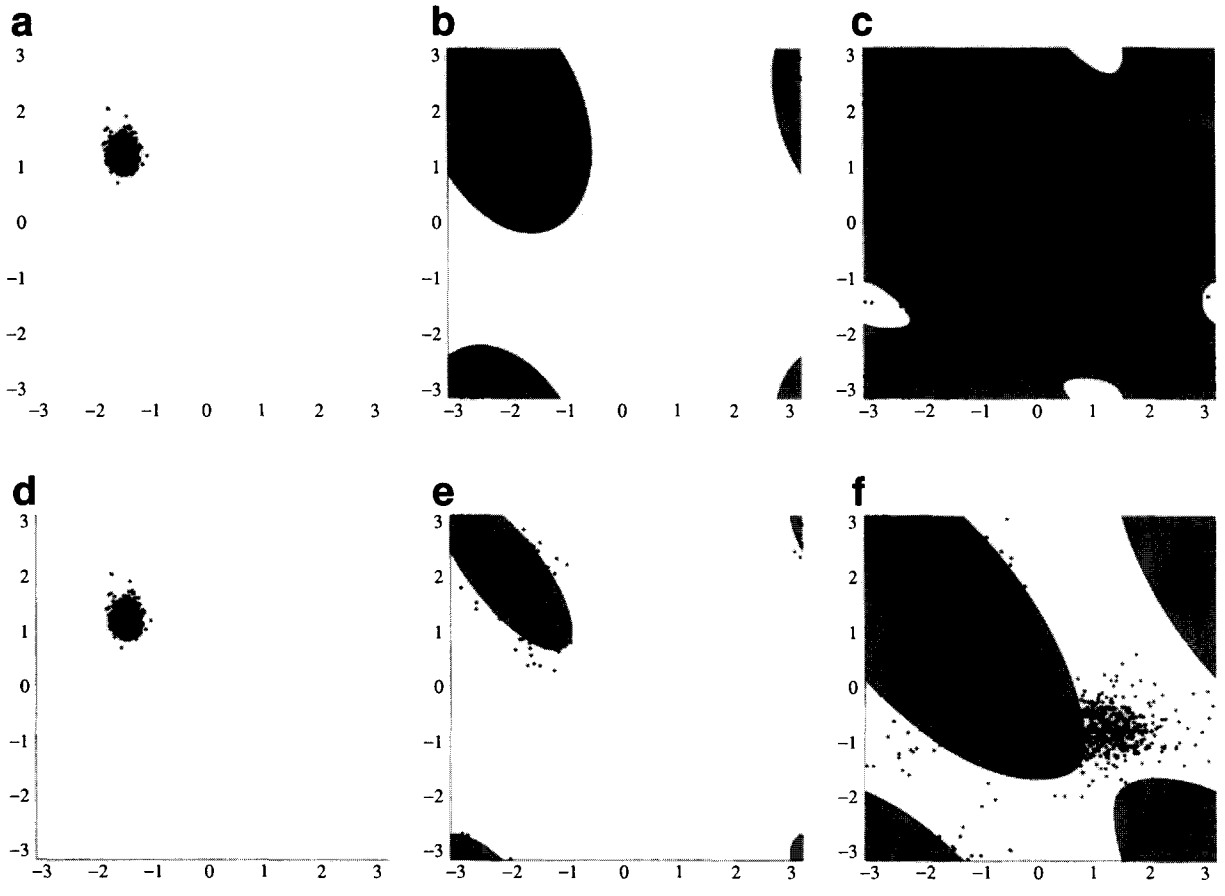

Figure 1 The ϕ–ψ scatter plots for terminally blocked valine at 100 K ((a) and (d)), 300 K ((b) and (e)) and 1100 K ((c) and (f)) showing the corresponding descriptions produced by the angular quasiharmonic ((a)–(c)) and multivariate wrapped-Gaussian ((d)–(f)) models. The shaded regions are bounded by the contours where $p(\phi,\psi)/p(\phi^0,\psi^0) = e^{-2}$ for $p \equiv p_{\angle}$ ((a)–(c)) or $p \equiv p_{\mathrm{wg}}$ ((d)–(f)). In contrast to the quasiharmonic method, the MWG method, the MWG method accurately models the data set even at high temperatures

because it ignores the periodicity of the ψ and ϕ variables both in calculating the mean angular position and in calculating the fluctuation tensor. (In this simple case, inspection suggests that a better approximtion could be obtained if the coordinate system were shifted by π in both the ϕ and ψ variables. However, in practical cases, such shifts must be determined algorithmically, not by visual inspection. The standard angular quasiharmonic method cannot do this. Furthermore, even a shift of origin will not result in accurate estimation of large angular fluctuations.) The failure is complete at 1100 K (*Figure 1c*), where the quasiharmonic approximation yields a model Gaussian distribution that is rather uniform over most of the region, even though significant angular dependences are visually obvious. A better method is needed.

Single-variable wrapped-Gaussian distribution

The von Mises distribution, $p_{\mathrm{vM}}(\theta;\kappa,\theta^0) \propto \exp[\kappa\cos\times(\theta-\theta^0)]$ is commonly used to describe angular statistics in a single variable[15]. While it gives good results, we do not know how to generalize it to the multivariate case. Instead, we generalize the single-variable *wrapped-Gaussian distribution*[15]. We begin by noting that the Gaussian distribution (1) is the solution of the diffusion equation in Cartesian coordinates with an anisotropic diffusion tensor. This motivates us to use the corresponding solution in periodic angular coordinates as a statistical model. Since the diffusion equation is linear, we can go from the solution on a line to the solution on a ring simply by 'wrapping' the probability in the region outside the interval $-\pi \le \theta < \pi$ onto that interval. *Figure 2* illustrates this for three one-dimensional cases: where $\xi \ll \pi$, $\xi = \sqrt{2}/2$ and $\xi > \pi$. The one-dimensional wrapped-Gaussian distribution is:

$$p_{wg}(\theta) = \frac{1}{\sqrt{2\pi}\xi}\sum_{m=-\infty}^{\infty}\exp[-(\theta-\theta^0-2\pi m)^2/2\xi^2] \quad (10)$$

It is proportional to the Jacobi ϑ function, which plays a prominent role in many areas of mathematics. It can also be expressed as a Fourier series:

$$p_{wg}(\theta) = \frac{1}{2\pi}\sum_{k=-\infty}^{\infty}\exp(-\xi^2k^2/2)\exp[-i\kappa(\theta-\theta^0)] \quad (11)$$

In the (low-temperature) limit where $\xi \ll \pi$ (e.g. as in the left panel of *Figure 2*), all the terms in representation (10) with $m \neq 0$ are negligible and p_{wg} approaches the standard Gaussian distribution:

$$p_{wg}(\theta) \approx \frac{1}{\sqrt{2\pi}\xi}\exp[-(\theta-\theta^0)^2/2\xi^2] \qquad \text{for } \xi \ll \pi \quad (12)$$

In the (high-temperature) limit where $\xi \ge \pi$, only the $k = 0, \pm 1$ terms in representation (11) remain significant and:

$$p_{wg}(\theta) \approx \frac{1}{2\pi}[1 + 2\exp(-\xi^2/2)\cos(\theta-\theta^0)] \qquad \text{for } \xi \ge \pi \quad (13)$$

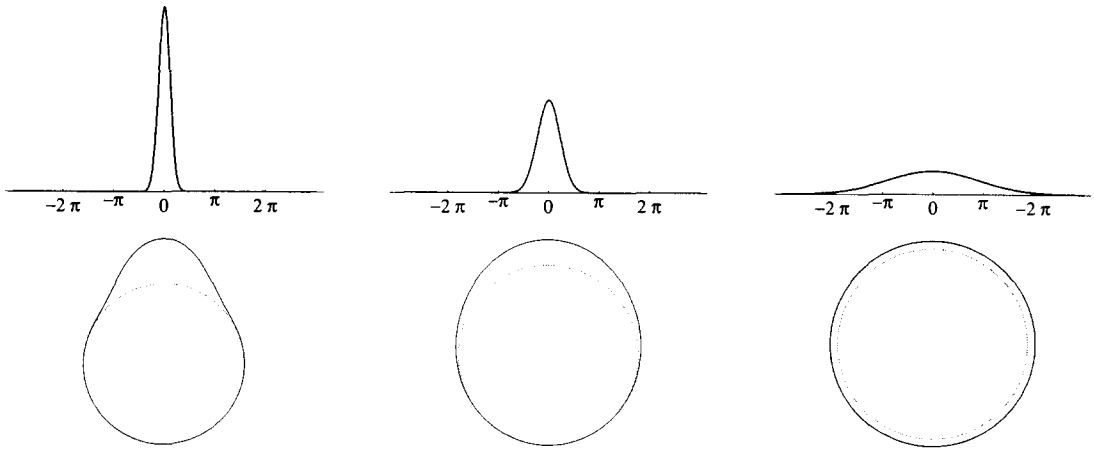

Figure 2 Three examples of one-dimensional wrapped-Gaussian distributions with (left to right) $\xi \ll \pi$, $\xi = \sqrt{2}/2$ and $\xi > \pi$

Like standard Gaussian distribution, the wrapped-Gaussian distribution is governed by two parameters, θ^0 and ξ, which describe its mean and fluctuation. However, unlike the scale-covariant Euclidean case (the scale covariance of the Gaussian distribution in Euclidean space permits the familiar transformation to scale-invariant z variables, $z \equiv (x-\mu)/\sigma$, which greatly simplifies analysis), the non-Euclidean angular space has a unique scale, π, set by its periodicity. When $\xi \ll \pi$ the curvature can be ignored and θ^0 and ξ can be well estimated using the quasiharmonic approximation (4) and (5), with appropriate substitutions. However, when $\xi \not\ll \pi$ a more sophisticated approach must be used. We proceed by analogy to the quasiharmonic equations (4) and (5), which determine the Gaussian parameters, by matching the expectation values of the first and second spatial moments of the data set to those predicted by the distribution. To allow for the periodicity of the space, we consider the moments, not of θ, but of $\exp(i\theta_\alpha)$, $\alpha = 1, \ldots, N^d$, where $\{\theta_\alpha\}$ is the data set of observed angles. We demand that:

$$\langle \exp(i\theta_j) \rangle = \int_{-\pi}^{\pi} e^{i\theta} p_{wg}(\theta)\, d\theta = \exp[-(\xi^2/2 - i\theta^0)] \quad (14)$$

This implies that:

$$\arg\langle e^{i\theta} \rangle = \theta^0 \quad (15)$$

$$-2\log|\langle e^{i\theta} \rangle| = \xi^2 \quad (16)$$

The positivity of the left-hand side of (16) is guaranteed, as required for a normalizable distribution, by the triangle inequality.

Equations (15) and (16) are similar to those used for determining the parameters of the von Mises distribution[15] and give reasonable approximations to one-dimensional angular distributions. They could also be used to describe fluctuations in multiple angular variables if the fluctuations were uncorrelated. The difficulty lies in developing method that can accommodate angular corrections.

Multivariate wrapped-Gaussian distribution

We generalize equation (10) to define the *multivariate wrapped-Gaussian (MWG) distribution*:

$$p_{\mathbf{wg}}(\boldsymbol{\theta}) = \det{}^{-1}(\sqrt{2\pi}\boldsymbol{\Xi}) \sum_{\boldsymbol{m} \in \boldsymbol{x}^N} \exp[-(\boldsymbol{\theta} - \boldsymbol{\theta}^0 - 2\pi\boldsymbol{m}) \times (2\boldsymbol{\Xi}^2)^{-1}(\boldsymbol{\theta} - \boldsymbol{\theta}^0 - 2\pi\boldsymbol{m})] \quad (17)$$

$$\int_{i\pi}^{\pi} p_{\mathbf{wg}}(\boldsymbol{\theta})\, d\boldsymbol{\theta} = 1 \quad (18)$$

where $\boldsymbol{\theta} \equiv \{\theta_j : j = 1, \ldots, N\}$ is a vector of angles in the N-dimensional angular space, $\boldsymbol{\theta}^0$ is the vector that specifies the centre of the distribution, $\boldsymbol{\Xi}^2$ is the fluctuation matrix, and $\boldsymbol{m}$ is a vector on the N-dimensional integer lattice, $\boldsymbol{Z}^N$. The sum is taken over the entire lattice. This is the solution for anisotropic diffusion on a multi-dimensional torus. It can also be expressed as a Fourier series:

$$p_{\mathbf{wg}}(\boldsymbol{\theta}) = (2\pi)^{-N} \sum_{\boldsymbol{k} \in \boldsymbol{Z}^N} \exp[-\boldsymbol{k} \cdot (\boldsymbol{\Xi}^2/2) \cdot \boldsymbol{k}] \times \exp[-i\boldsymbol{k} \cdot (\boldsymbol{\theta} - \boldsymbol{\theta}^0)] \quad (19)$$

As in the one-dimensional case, we might hope to determine the elements of $\boldsymbol{\theta}^0$ and $\boldsymbol{\Xi}^2$ by matching the expectation values of exponential moments calculated from the data set to those predicted by the MWG distribution. Thus we generalize (14) to:

$$\langle e^{i\boldsymbol{\eta} \cdot \boldsymbol{\theta}} \rangle = \int e^{i\boldsymbol{\eta} \cdot \boldsymbol{\theta}} p_{\mathbf{wg}}(\boldsymbol{\theta})\, d\boldsymbol{\theta} = \exp[-\eta \cdot (\boldsymbol{\Xi}^2/2) \cdot \boldsymbol{\eta}] \qquad \boldsymbol{\eta} \in \boldsymbol{Z}^N \quad (20)$$

This yields relations that depend on $\boldsymbol{\theta}^0$:

$$\arg\langle e^{i\boldsymbol{\eta} \cdot \boldsymbol{\theta}} \rangle = \boldsymbol{\eta} \cdot \boldsymbol{\theta}^0 \qquad \boldsymbol{\eta} \in \mathbf{Z}^N \quad (21)$$

and on $\boldsymbol{\theta}^2$:

$$-2\log|\langle e^{i\boldsymbol{\eta} \cdot \boldsymbol{\theta}} \rangle| = \boldsymbol{\eta} \cdot \boldsymbol{\Xi}^2 \cdot \boldsymbol{\eta} \qquad \boldsymbol{\eta} \in \boldsymbol{Z}^N \quad (22)$$

Equations (21) and (22) cannot be satisfied for all $\boldsymbol{\eta}$ since we only have $N + N(N+1)/2$ variable parameters. Following the one-dimensional procedure, it seems reasonable to fix the N elements of $\boldsymbol{\theta}^0$ by evaluating (21) for the first-order on-axis moments of the θ_j:

$$\arg\langle \exp(i\theta_j) \rangle = \theta_j^0 \quad (23)$$

In the same vein, we would like to fix the N diagonal elements of $\boldsymbol{\Xi}^2$ by evaluating (22) for the first-order on-axis moments:

$$-2\log|\langle \exp(i\theta_j) \rangle| = \Xi_{jj}^2 \qquad \text{[not used]} \quad (24)$$

and to fix its $N(N-1)/2$ off-diagonal elements by using (22) with combinations of the lowest-order

off-axis moments $\langle \exp[i(\theta_\xi \pm \theta_k)]\rangle$:

$$-\frac{1}{2}\log\frac{|\langle \exp[i(\theta_j - \theta_k)]\rangle|}{|\langle \exp[i(\theta_j + \theta_k)]\rangle|} = \Xi^2_{jk} \qquad \text{[not used]} \qquad (25)$$

However, (24) and (25) cannot be used to fix Ξ^2 because they do not guarantee that the calculated Ξ^2 will be positive semidefinite for every data set. (These equations will yield positive semidefinite Ξ^2 if the data set is derived from a wrapped-Gaussina distribution and contains a very large number of points.) IF Ξ^2 is indefinite (i.e. has negative eigenvalues) the distribution generated by (17) will not be normalizable. We find that this frequently occurs with typical data sets, particularly when they are small.

An alternative approach is suggested by the recognition that the $\sin(\theta_j - \theta_j^0)$ provide normalized periodic variables that are somewhat analogous to the R_j variables used in Cartesian space. Roughly following the Cartesian procedure, we might attempt to fix all the elements of Ξ^2 by the $N(N+1)/2$ relations (derived from (21) and (22)):

$$\frac{\langle \sin(\theta_j - \theta_j^0)\sin(\theta_k - \theta_k^0)\rangle}{\langle \exp[i(\theta_j - \theta_j^0)]\rangle\langle \exp[i(\theta_k - \theta_k^0)]\rangle} = \sinh(\Xi^2_{jk}) \qquad \text{[not used]} \qquad (26)$$

where $\boldsymbol{\theta}^0$ is fixed by (21).

The matrix on the left-hand side of (26) is the expectation value of an outer product ((defining $s_j \equiv \sin(\theta_j - \theta_j^0)/\langle \exp[i(\theta_j - \theta_j^0)]\rangle$, the left-hand side of (26) can be expressed as $\langle s_j s_k\rangle$), and thus is guaranteed to be positive semidefinite. Unfortunately, the inverse hyperbolic operation required to calculate Ξ^2 from this matrix will not necessarily preserve this property, and this method also fails for some data sets.

To remedy this problem we seek a way of calculating an *approximate* Ξ^2 that is guaranteed to be positive semidefinite and, moreover, that can be calculated efficiently. We are guided by three requirements:

1. Ξ^2 must have only positive eigenvalues.
2. When the fluctuations are small (i.e. when the eigenvalues of Ξ^2 are $\ll \pi^2$), Ξ^2 should approximately satisfy (24), (25) and (26).
3. When the angular fluctuations are uncorrelated, the diagonal elements of Ξ^2 should be determined by (24), which is equivalent to using (16) separately for each variable.

We begin by modifying (26) to force it to preserve the positive semidefinite property. For this purpose, we make the approximation of replacing the matrix of hyperbolic sines with the hyperbolic sine of the matrix:

$$\frac{\langle \sin(\theta_j - \theta_j^0)\sin(\theta_k - \theta_k^0)\rangle}{\langle \exp[i(\theta_j - \theta_j^0)]\rangle\langle \exp[i(\theta_k - \theta_k^0)]\rangle} = (\sinh \Xi^2)_{jk} \qquad \text{[not used]} \qquad (27)$$

This replacement, while not accurate in general, is accurate when Ξ^2 is small or almost diagonal. Because the inverse hyperbolic sine is monotonic and positive for positive argument, x^2 determined by (27) is guaranteed to be positive semidefinite and, thus, to satisfy the first requirement. The second requirement is also satisfied because, for small fluctuations, Ξ^2 will be small and $\sinh \Xi^2$ can be approximated by x^2 in (27). The resultant expression is equivalent to (24), (25) and (26) when similar small-fluctuation linear approximations to their log and sinh terms are made. However, the third requirement is not satisfied when the fluctuations are uncorrelated but large since the left-hand side of (27) depends on the second-order moments $\langle \exp[i2(\theta_j - \theta_j^0)]\rangle$ while the left-hand side of (24) depends only on the first-order moments $\langle \exp[i(\theta_j - \theta_j^0)]\rangle$.

To obtain an approximation that satisfies all three requirements, we first note that the expectation value in the left-hand side of (24) can be rewritten as:

$$|\langle \exp(i\theta_j)\rangle| = 1 - 2\langle \sin^2[\tfrac{1}{2}(\theta_j - \theta_j^0)]\rangle \qquad (28)$$

This suggests that we fix Ξ^2 by matching expectation values of:

$$S_{jk} \equiv \sin[\tfrac{1}{2}(\theta_j - \theta_j^0)]\sin[\tfrac{1}{2}(\theta_k - \theta_k^0)]$$

Matching $\langle S_{jk}\rangle$ to the corresponding expectation predicted by the MWG distribution requires the evaluation of (20) for half-integer-order moments. Although equation (20) is only exact for $\boldsymbol{\eta}$ on the integer lattice, we expect (and numerical tests verify) that it provides an interpolating function that is approximately correct at the half-integer values needed to evaluate $\int S_{jk}(\boldsymbol{\theta}) p_{\mathbf{wg}}(\boldsymbol{\theta})\, d\boldsymbol{\theta}$. Using this approximation and the same line of reasoning that led to (27), we get:

$$\frac{\langle \sin[\tfrac{1}{2}(\theta_j - \theta_j^0)]\sin[\tfrac{1}{2}(\theta_k - \theta_k^0)]\rangle}{\langle \exp[i(\theta_j - \theta_j^0)]\rangle^{1/4}\langle \exp[i(\theta_k - \theta_k^0)]\rangle^{1/4}} = (\sinh \Xi^2/4)_{jk} \qquad (29)$$

Equation (29) with (23) satisfies all three requirements and thus is a candidate for calculating the $\boldsymbol{\theta}^0$ and Ξ^2 parameters to match an MWG distribution to an arbitrary data set. The required expectation values can be directly calculated and the hyperbolic sine can be inverted either in the eigenvector basis or by using a few terms of a matrix power series. Since the derivation of the parameter-fitting equations is heuristic, we have no proof or measure of its relative accuracy. However, empirically we have found that it is a robust method that accurately models the correlated distributions that are encountered in the analysis of polymer torsion-angle distributions. (Numerical tests show that, as expected, (29) proves a better fit than (27).) This is demonstrated in the next section for two-dimensional cases that can be graphically analysed.

APPLICATION OF THE MWG METHOD TO TWO-ANGLE PROBLEMS

We can test the utility of the MWG distribution (17) with the parameter-fixing conditions (23) and (29) by re-analysing the terminally blocked valine data set. This gives the e^{-2} regions shown in *Figure 1d–f*. In contrast to the quasiharmonic method, the MWG method approximates the distribution well even at the highest temperature.

To demonstrate further the power of the method, we consider the collection of scatter plots for two-angle data sets having different amounts of fluctuation and correlation that are displayed in *Figure 3*. The e^{-2} regions for the angular quasiharmonic fits to these distributions

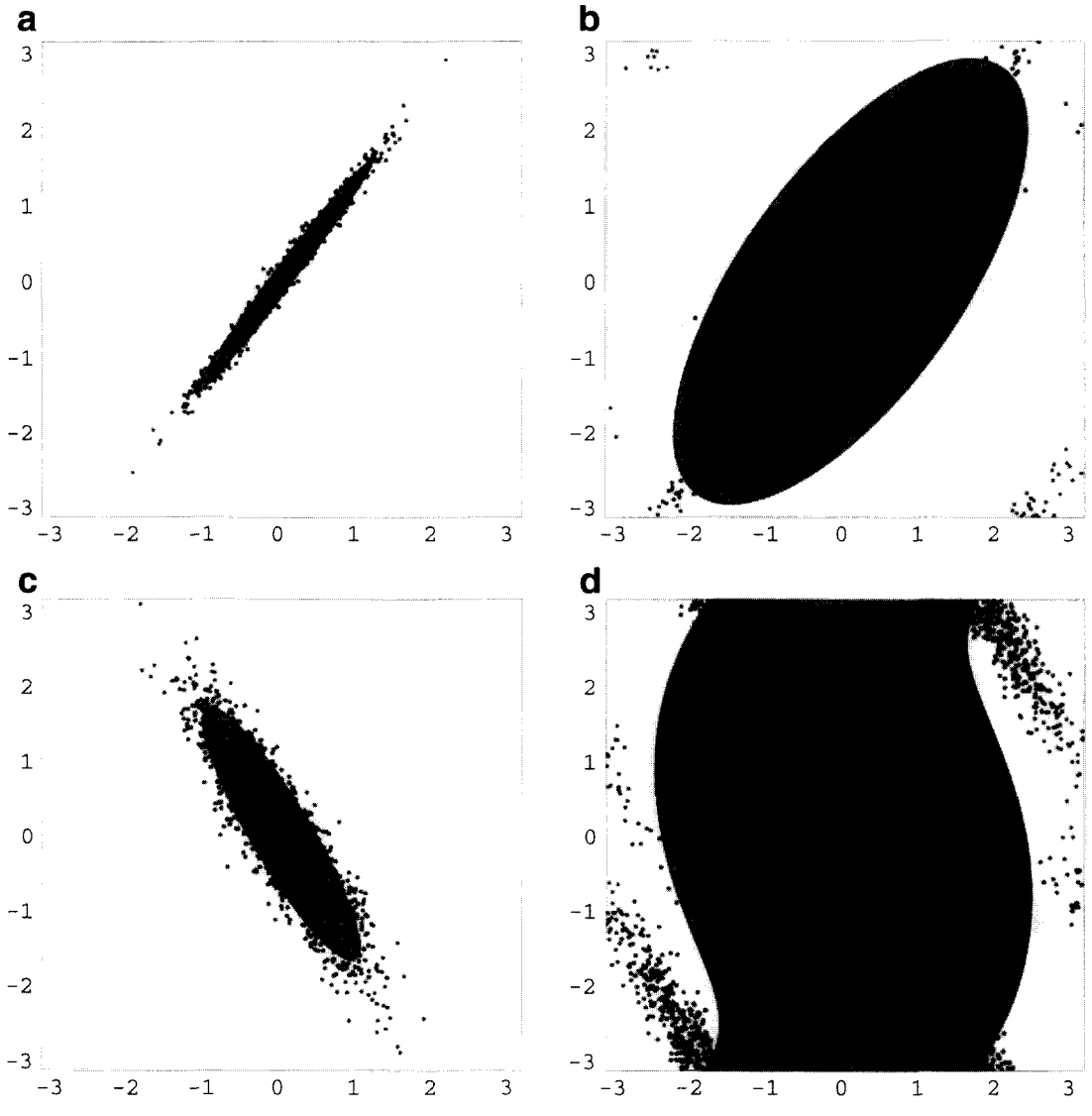

Figure 3 Scatter plots of two-dimensional data sets having various amounts of correlation and fluctuation. Panels (a)–(d) show the e^{-2} regions generated by the angular quasiharmonic distribution; panels (e)–(h) show the corresponding regions generated by the MWG distribution

are shown in *Figures 3a–d*; the MWG regions are shown in *Figures 3e–h*. The properties noted in the terminally blocked valine example are observed in these cases as well. In general, the quasiharmonic method is only accurate for small fluctuations and tends to overestimate the fluctuation tensor when fluctuations are large. This occurs because the quasiharmonic method only allows for (on-axis) periodicity of the individual angles and not for the (off-axis) periodicity that results from correlated angular fluctuations. Thus, for example, it is unable to recognize that the data points in the lower right and upper left corners of *Figure 3b* are periodically related to the main lobe and yields an incorrectly large estimate of the width of the distribution. In contrast, the periodic MWG method recognizes these off-axis periodicities and is accurate even with very distributions displaying large correlated angular fluctuations. The example shown in *Figure 3h* is particularly remarkable: it shows that the MWG can accurately model a correlated distribution that completely wraps around the interval $(-\pi, \pi)$ in angle space.

APPLICATION OF THE MWG METHOD TO PEPTIDES

The ability of the method to characterize fluctuations in higher-dimensionality problems was tested by analysing the thermal motions of the peptide Met-enkephalin (H_2N-Tyr-Gly-Gly-Phe-Met-COOH) *in vacuo*. This peptide contains 75 atoms and its conformation is described by 24 torsion angles. Peptide bonds were assumed to be rigid, thereby reducing the number of angles to 19. Thermal motions of the peptide were computationally sampled using the ECCEP3[17] empirical potential with Metropolis Monte Carlo[18] importance sampling. Fluctuations were examined about the previously calculated[19] lowest-energy conformation using a limited data set of 1000 conformations. Our purpose was only to compare the characterization of the same data set by the angular quasiharmonic and MWG methods, so no attempt was made to ensure that the entire conformation space of Met-enkephalin was sampled.

Both highly correlated and uncorrelated eigenmodes were observed. Four representative eigenvectors

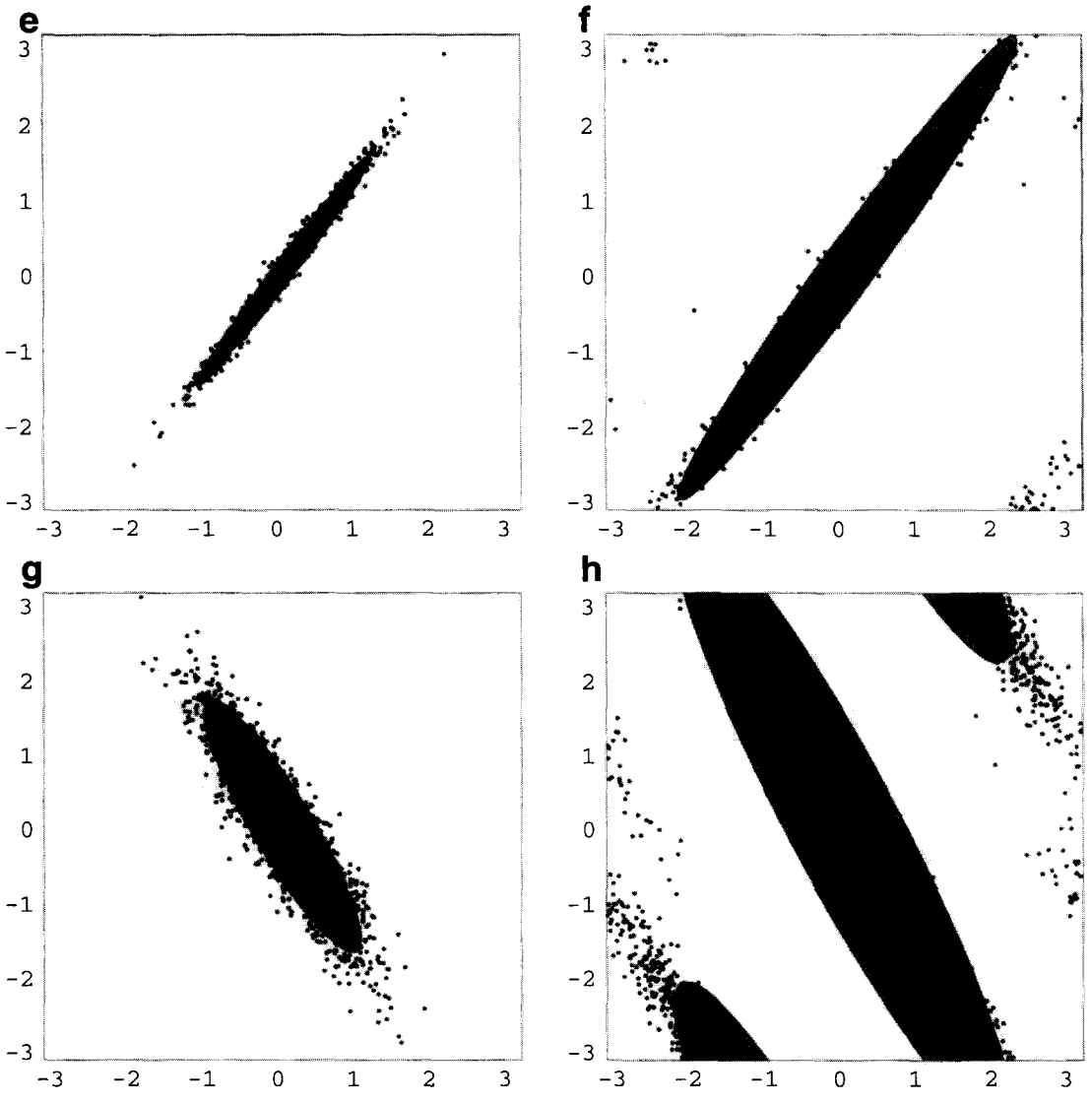

Figure 3 Continued

(out of the total of 19) are shown in *Figure 4*. The mode shown in the upper left panel involves the uncorrelated motion of a single torsion angle at the amino end of the chain. In contrast, the mode shown in the lower right panel involves the correlated motion of several angles in the centre of the protein.

The root-mean-square fluctuations of many of the torsion angles were in the range of 20–60°. Some were as large as 120°. This is reflected in the relatively large magnitudes of the eigenvector components shown in *Figure 4*. Because these fluctuations are large, they could not have been well modelled by using the harmonic or quasiharmonic approximations in either Cartesian or torsion-angle space. To demonstrate the magnitude of the errors that can be introduced by inappropriate use of the quasiharmonic approximation, we compared the amplitudes of all the eigenmodes as calculated by the MWG and torsion-angle quasiharmonic methods. *Figure 5* displays the (ordered) squared amplitudes of all 19 eigenmodes at low (300 K) and high (1350 K) temperatures. As shown in *Figures 5c* and *5d*, the two methods give very different results at high temperature. The MWG analysis clearly indicates that mean-square

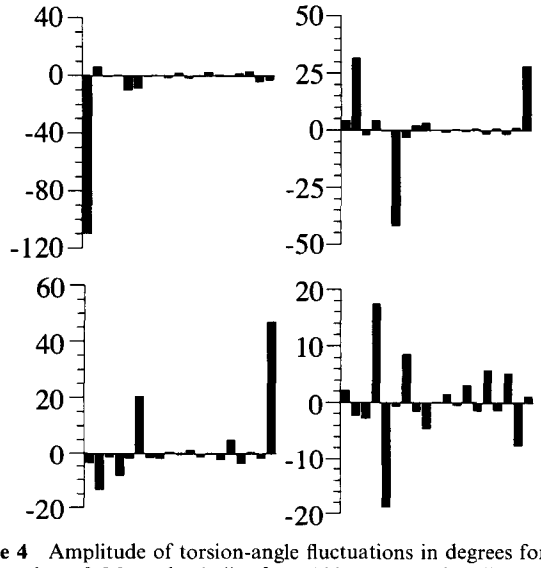

Figure 4 Amplitude of torsion-angle fluctuations in degrees for four eigenmodes of Met-enkephalin for 1000 computationally sampled conformations at 300 K

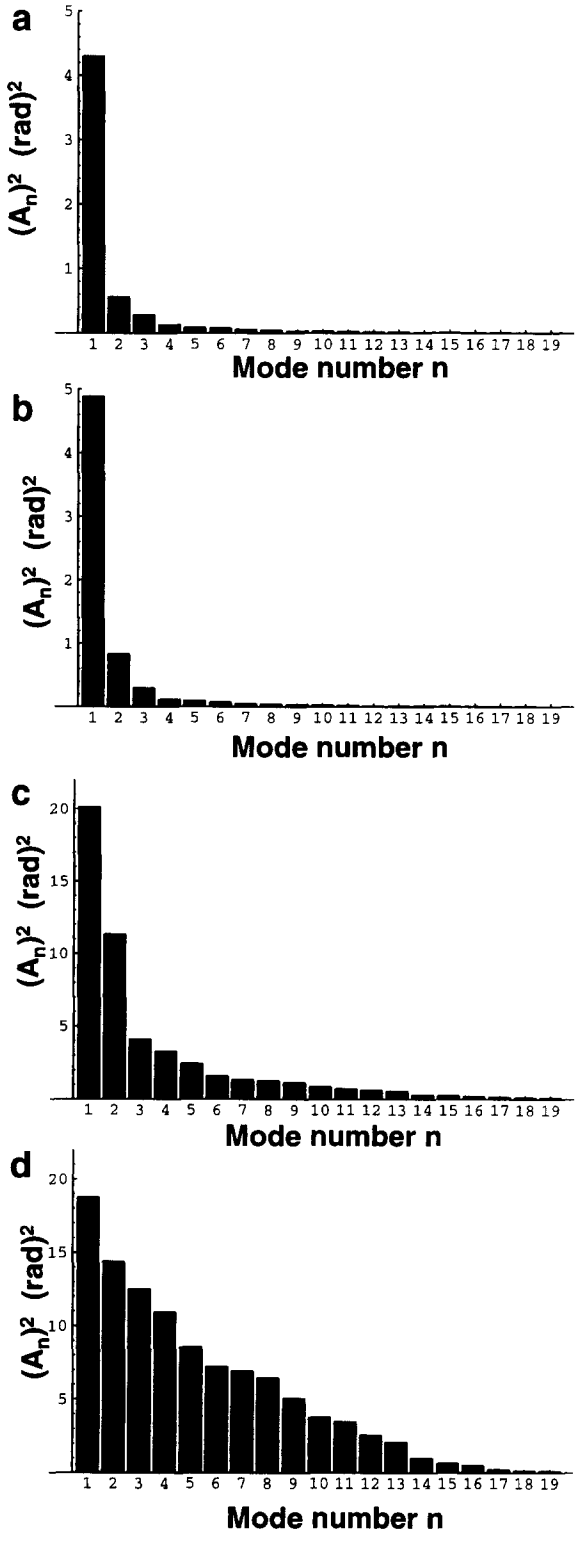

Figure 5 Eigenvalues of the angular fluctuation matrix Ξ^2 calculated for Met-enkephalin from Monte Carlo data sets generated at 300 K ((a) and (b)) and 1350 K ((c) and (d)) using either the MWG ((a) and (c)) or angular quasiharmonic ((b) and (d)) models

angular fluctuations are dominated by only a few modes. The extent of this domination is not fully represented by the inaccurate quasiharmonic description, which, as in the example of terminally blocked valine, tends to overestimate the angular fluctuations. As expected, the results from the two methods are much more similar at low temperature (cf. *Figure 5a* and *5b*). However, even here, while there are no significant qualitative differences, the quasiharmonic method overestimates the square amplitude of mode 2 by ~40%. Further processing of the data would be required to relate these modes to observable thermodynamic properties, but these comparisons suffice to indicate the advantages of the MWG method.

CONCLUSION

The MWG distribution is useful for describing distributions of polymer conformations in internal angular coordinates. It is equivalent to previously described methods when describing distributions having only small fluctuations, but provides a significantly more accurate description when fluctuations are large. This is particularly important with small polymers or when large polymers are simulated at high temperatures to search large areas of conformation space.

The terminally blocked valine example showed how the conventional angular quasiharmonic method tends to overestimate the extent of correlated fluctuations to the point where significant information is lost. This tendency towards overestimation probably explains the fact that the quasiharmonic method reported larger fluctuations for Met-enkephalin relative to the MWG method. This artifact obscures the extent to which fluctuations in the Met-enkephalin data set were concentrated in a small number of dominant modes.

The superior accuracy of the MWG method relative to the angular quasiharmonic method derives from its respect of the periodicity of the angle space. Both methods are integral methods and, in contrast to the differential harmonic method, there is no need for the underlying potential to be roughly harmonic for the MWG method to be applicable. Both methods characterize distributions in terms of their first and second moments, and even highly anharmonic multinodal distributions (e.g. like that shown in *Figures 1e* and *1f*) can be well described. In Cartesian space, it is possible to prove that the eigenvectors of the quasiharmonic Λ^2 calculated using equation (5) describe the set of collective motions that best represents the largest fluctuations in a least-squares sense (e.g. see Garcia[20]). While we cannot construct a formal proof for our approximate method, it seems evident that the MWG method provides an analogous integral characterization in angle space. Thus, we expect that the collective motions corresponding to the eigenvectors of Ξ^2 similarly describe the dominant motions of polymers.

An important application of the MWG method will be in calculating anisotropic transition functions for Monte Carlo sampling of proteins in torsion-angle space. We expect that its increased accuracy will result in improved efficiency relative to direct application of the Vanderbilt and Louie method[9], which uses a Gaussian distribution like (1) for the transition function, to proteins[10,11]. In this regard, we note that random vectors having the MWG distribution density can be easily generated. This accomplished by generating random vectors in an 'unwrapped' unbounded θ space with the Gaussian distribution (1) with substitutions (8) and (9) and simply mapping each vector coordinate back into the interval $(-\pi, \pi)$.

Determination of the MWG fluctuation tensor from the data set is a non-trivial problem because of the difficulties encountered in ensuring that it be positive

semidefinite. The approximate method we have developed works very well in practice. However, it is deficient on two accounts. First, it tends to overestimate the width of extremely elongated, correlated distributions (such as that shown in *Figure 3h*) that 'wrap' around the angle space in an off-axis direction. Secondly, when a data set is numerically generated using the wrapped-Gaussian distribution with a non-diagonal fluctuation tensor Ξ^2 as a random kernel, equation (29) does not exactly recover the original Ξ^2. This is true even when the size of the data set, N^d, is taken to infinity. In our experience neither of these deficiencies is important for polymer calculations: extreme distributions like that shown in *Figure 3* are not encountered and (29) provides a good approximation to the original Ξ^2 for MWG distributions within the observed range. Nonetheless, a method for calculating Ξ^2 that addressed these points could further improve accuracy to some extent and extend the range of applicability of the MWG approach. Formal implicit equations (e.g. a maximum likelihood estimator) that do recover the original Ξ^2 in the limit $N^d \to \infty$ can be derived. But the iterative computations needed for solution are prohibitively expensive for the high-dimensionality distributions needed to characterize large polymers. Further work is needed.

REFERENCES

1 Brooks, B. and Karplus, M. *Proc. Natl Acad. Sci. USA* 1993, **80**, 6571
2 Karplus, M. and Kushick, J. N. *Macromolecules* 1981, **14**, 325
3 Armadei, A., Linssen, A. B. and Berendsen, H. J. C. *Proteins: Struct. Funct. Genet.* 1993, **17**, 412
4 Horiuchi, T. and Gō, N. *Proteins: Struct. Funct. Genet.* 1991, **10**, 106
5 Gō, N., Noguti, T. and Nishikawa, T. *Proc. Natl Acad. Sci. USA* 1985, **80**, 3696
6 Wille, L. *Nature* 1986, **324**, 46
7 McCammon, J. A. *Nature* 1976, **262**, 325
8 Noguti, T. and Gō, N. *Biopolymers* 1985, **24**, 527
9 Vanderbilt, D. Louie, S. G. *J. Comput. Phys.* 1984, **56**, 259
10 Shin, J. K. and Jhon, M. S. *Biopolymers* 1991, **31**, 177
11 Yoon, J. H., Shin, J. K. and Jhon, M. S. *J. Comput. Chem.* 1995, **16**, 478
12 Church, B., Oresic, M. and Shalloway, D. in 'DIMACS Series in Discrete Mathematics and Theoretical Computer Science' (Eds P. Pardalos, D. Shalloway and G. Xue), Americal Mathematical Society, Providence, RI, 1995 (in press)
13 Jensen, L. H. in 'Methods in Enzymology' (Eds H. W. Wyckoff, C. Hirs and S. Timasheff), Vol. 115, Academic Press, New York, 1985, Ch. 16, pp. 227–234
14 Brunger, A., Krukowski, A. and Erickson, J. *Acta Crystallogr. (A)* 1990, **46**, 585
15 Mardia, K. V. 'Statistics of Directional Data', Academic Press, London, 1972, pp. 55–56
16 Elber, R. *et al.* in 'Statistical Mechanics, Protein Structure, and Protein Substrate Interactions' (Ed. S. Doniach), Plenum Press, New York, 1994, pp. 165–191
17 Nemethy, G. *et al. J. Chem. Phys.* 1992, **96**, 6472
18 Metropolis, N., Metropolis, A. W., Rosenbluth, M., Teller, A. H. and Teller, E. *J. Chem. Phys.* 1953, **21**, 1087
19 Li, Z. and Scheraga, H. A. *Proc. Natl Acad. Sci. USA* 1987, **84**, 6611
20 Garcia, A. E. *Phys. Rev. Lett.* 1992, **68**, 2696