Least constraint approach to the extraction of internal motions from molecular dynamics trajectories of flexible macromolecules

Guillaume Chevrot,[1,2] Paolo Calligari,[3] Konrad Hinsen,[1,2] and Gerald R. Kneller[1,2,4,a)]
[1]*Centre de Biophys. Moléculaire, CNRS; Rue Charles Sadron, 45071 Orléans, France*
[2]*Synchrotron Soleil; L'Orme de Merisiers, 91192 Gif-sur-Yvette, France*
[3]*Département de Chimie, associé au CNRS, Ecole Normale Supérieure, 24, rue Lhomond, 75231 Paris Cedex 05, France*
[4]*Université d'Orléans; Chateau de la Source-Av. du Parc Floral, 45067 Orléans, France*

(Received 23 May 2011; accepted 29 July 2011; published online 24 August 2011)

We propose a rigorous method for removing rigid-body motions from a given molecular dynamics trajectory of a flexible macromolecule. The method becomes exact in the limit of an infinitesimally small sampling step for the input trajectory. In a recent paper [G. Kneller, J. Chem. Phys. **128**, 194101 (2008)], one of us showed that virtual internal atomic displacements for small time increments can be derived from Gauss' principle of least constraint, which leads to a rotational superposition problem for the atomic coordinates in two consecutive time frames of the input trajectory. Here, we demonstrate that the accumulation of these displacements in a molecular-fixed frame, which evolves in time according to the virtual rigid-body motions, leads to the desired trajectory for internal motions. The atomic coordinates in the input and output trajectory are related by a roto-translation, which guarantees that the internal energy of the molecule is left invariant. We present a convenient implementation of our method, in which the accumulation of the internal displacements is performed implicitly. Two numerical examples illustrate the difference to the classical approach for removing macromolecular rigid-body motions, which consists of aligning its configurations in the input trajectory with a fixed reference structure. © *2011 American Institute of Physics*. [doi:10.1063/1.3626275]

I. INTRODUCTION

The construction of trajectories representing the internal motions of a flexible macromolecule from a molecular dynamics (MD) trajectory is a recurrent task in biomolecular simulations. Such trajectories are particularly useful for the analysis of spectroscopic experiments probing the total dynamics of the molecule, but where the internal dynamics is of particular interest. Examples are combined experimental and simulation studies of proteins in solution by nuclear magnetic resonance (NMR) relaxation spectroscopy[1–4] and by quasi-elastic neutron scattering.[5]

The common strategy to extract the internal motions from a given MD trajectory is to align the snapshots of the protein with a common reference structure. The coordinate frame associated with that structure can be considered as the Eckart frame of the molecule, referring to the early work of Eckart on the theory of spectroscopic experiments on small molecules in the gas phase.[6] Eckart considered only small vibrations as possible internal motions and assumed the internal energy of the molecule to be a quadratic function of the atomic displacements. Observing that rigid-body displacements do not alter the internal potential energy of the molecule, he constructed conditions for the atomic displacements corresponding to internal motions of the molecule which exclude rigid-body displacements, and he also described a construction of a molecule-fixed frame describing the global motions of the molecule.

Shortly later Eckart's approach was generalized by Sayvetz to linear and "anomalous" molecules, which were excluded in Eckart's work.[7] Here "anomalous" refers to molecules which have internal motions of large amplitudes, such as rotations of methyl groups, which cannot be treated within the approximation of small vibrations. Motivated by the observation by Kudin and Dymarsky that the Eckart axis conditions are closely related to the problem of an optimal rotational superposition of molecular structures,[8] one of us (G.R.K.) has recently shown that virtual atomic displacements describing the internal motions in arbitrary macromolecules can be derived from Gauss' principle of least constraint,[9] which leads indeed to a rotational superposition problem of molecular structures.[10] The adjective "virtual" indicates that the internal displacements are reconstructed from a "real" trajectory including global and internal motions. At each time, one considers the motion that a virtual rigid molecule would have performed within an infinitesimal time interval, given the same initial atomic positions, velocities, and forces. The internal atomic displacements are then the differences between the real displacements and the displacements due to an infinitesimal virtual rigid-body motion. According to Gauss' formulation of mechanics, the internal displacements are minimized in a least square sense and fulfill automatically the Eckart conditions.

The aim of this paper is to discuss the construction of trajectories for the internal dynamics of macromolecules from the time-local virtual displacements obtained from Gauss' principle. A straightforward proposition has been made in Ref. 10, suggesting that the (almost) infinitesimal

a)Author to whom correspondence should be addressed. Electronic mail: gerald.kneller@cnrs.orleans.fr.

displacements obtained from an equidistantly sampled molecular dynamics trajectory could be simply accumulated in the laboratory frame. Here, it is shown that the accumulation should instead be done in a common molecular-fixed frame, which can be considered as the Eckart frame of the molecule and which can be formally constructed by an accumulation of the infinitesimal virtual rigid-body motions obtained from the least constraint principle. The theoretical part of the paper is presented in Sec. II, starting with a short review of the essential points of Ref. 10. In Sec. III, practical aspects concerning the explicit construction of the molecule-fixed frame and the corresponding internal-motion trajectories are discussed. Two applications are presented in Sec. IV to illustrate the difference between the traditional method of removing global motions and the one we propose. The paper is concluded by a short discussion of the results.

II. THEORY

A. Virtual internal displacements

This section gives a short review of the relation between the constraints for internal atomic motions in flexible macromolecules formulated by Eckart[6] and Gauss' principle of least constraint,[9,11–13] which has been developed in Ref. 10. According to Gauss the motion of a mechanical system under constraints can be derived from a local minimum principle. Considering N point-like particles with given positions $\mathbf{x}_\alpha$ at time t and masses m_α, the principle can be formulated as

$$\xi = \frac{1}{2}\sum_{\alpha=1}^{N} m_\alpha \left(\mathbf{x}_\alpha(t+\delta t) - \mathbf{x}_\alpha^{(c)}(t+\delta t)\right)^2 = \text{Min}, \tag{1}$$

where $\mathbf{x}_\alpha^{(c)}(t+\delta t)$ are the constrained positions at time $t+\delta t$. The latter are thus obtained from a least squares fit to the unconstrained positions at time $t+\delta t$. The time increment δt must be small enough to permit the approximation $\mathbf{x}_\alpha(t+\delta t) \approx \mathbf{x}_\alpha(t) + \delta t\, \dot{\mathbf{x}}_\alpha(t) + \delta t^2\, \mathbf{F}_\alpha(t)/(2m_\alpha)$, where the dot denotes a derivative with respect to time and $\mathbf{F}_\alpha(t)$ is the force acting on atom α. Suppose now that the coordinates $\mathbf{x}_\alpha(t) = \mathbf{X}(t) + \mathbf{r}_\alpha(t)$ define the atomic positions in a "virtual rigid molecule" at time t, where $\mathbf{r}_\alpha$ are the relative positions with respect to a common rotation center, $\mathbf{X}$. Within the time span δt, the atomic positions in the virtual rigid molecule will evolve to

$$\mathbf{x}_\alpha^{(c)}(t+\delta t) = \mathbf{X}(t) + \mathbf{r}_\alpha(t) + \boldsymbol{\delta\tau} + \boldsymbol{\delta\phi} \wedge \mathbf{r}_\alpha(t), \tag{2}$$

where $\boldsymbol{\delta\tau}$ and $\boldsymbol{\delta\phi}$ define, respectively, a small translation and rotation of the virtual rigid molecule. Here, the symbol "$\wedge$" denotes a vector product. The vector $\boldsymbol{\delta\phi}$ points into the direction of the rotation axis and its modulus is the rotation angle. The differences between the real atomic positions, $\mathbf{x}_\alpha(t+\delta t)$, and the virtual atomic positions, $\mathbf{x}_\alpha^{(c)}(t+\delta t)$, define the atomic displacements due to the *internal* motions of the molecule,

$$\boldsymbol{\delta}\mathbf{u}_\alpha(t+\delta t) = \mathbf{x}_\alpha(t+\delta t) - \mathbf{x}_\alpha^{(c)}(t+\delta t). \tag{3}$$

The virtual rigid-body displacement is performed according to the minimum principle (1). Inserting here the form (2) for the constrained positions at time $t+\delta t$ leads to the necessary conditions[10]

$$\frac{\partial \xi}{\partial \boldsymbol{\delta\tau}} = \sum_{\alpha=1}^{N} m_\alpha \boldsymbol{\delta}\mathbf{u}_\alpha(t+\delta t) = \mathbf{0}, \tag{4}$$

$$\frac{\partial \xi}{\partial \boldsymbol{\delta\phi}} = \sum_{\alpha=1}^{N} m_\alpha \mathbf{r}_\alpha(t) \wedge \boldsymbol{\delta}\mathbf{u}_\alpha(t+\delta t) = \mathbf{0}, \tag{5}$$

which are the Eckart conditions for the internal atomic displacements.[6] The translational Eckart condition (4) and the rotational Eckart condition (5) constitute a system of linear equations for the components of $\boldsymbol{\delta\tau}$ and $\boldsymbol{\delta\phi}$, which can be decoupled if $\mathbf{X}$ is taken to be the center of mass,

$$\mathbf{X} = \frac{1}{M}\sum_{\alpha} m_\alpha \mathbf{x}_\alpha, \tag{6}$$

where M is the total mass of the molecule. In this case, one obtains

$$\boldsymbol{\delta\tau} = \mathbf{X}(t+\delta t) - \mathbf{X}(t) \tag{7}$$

and the rotation vector is given by

$$\boldsymbol{\delta\phi} = \boldsymbol{\theta}(t)^{-1}\mathbf{L}(t)\delta t, \tag{8}$$

where $\mathbf{L} = \sum_\alpha m_\alpha \mathbf{r}_\alpha \wedge \dot{\mathbf{r}}_\alpha$ is the angular momentum and $\boldsymbol{\theta}$ is the tensor of inertia, with components $\theta_{ij} = \sum_{\alpha=1}^{N} m_\alpha(|\mathbf{r}_\alpha|^2\delta_{ij} - r_{\alpha,i}r_{\alpha,j})$. Both quantities are referred to the laboratory frame.

B. Trajectories of internal motions

We consider a trajectory $\mathbf{x}_\alpha(n\,\delta t) \equiv \mathbf{x}_\alpha(n)$, where δt is a fixed small time increment justifying the approximations

$$\boldsymbol{\delta\tau}(n) \approx \delta t\, \dot{\mathbf{X}}(n), \tag{9}$$

$$\boldsymbol{\delta\phi}(n) \approx \delta t\, \boldsymbol{\omega}(n), \tag{10}$$

such that the constrained positions at $t = n\delta t$ are approximated by

$$\mathbf{x}_\alpha^{(c)}(n) \approx \mathbf{x}_\alpha(n-1) + \delta t\, \dot{\mathbf{X}}(n-1) + \delta t\, \boldsymbol{\omega}(n-1) \wedge \mathbf{r}_\alpha(n-1).$$

Using that $(\mathbf{x}_\alpha(n) - \mathbf{x}_\alpha(n-1))/\delta t \approx \dot{\mathbf{x}}_\alpha(n-1)$, one obtains

$$\begin{aligned}\boldsymbol{\delta}\mathbf{u}_\alpha(n) &= \mathbf{x}_\alpha(n) - \mathbf{x}_\alpha^{(c)}(n)\\ &\approx \delta t\{\dot{\mathbf{x}}_\alpha(n-1) - \dot{\mathbf{X}}(n-1) - \boldsymbol{\omega}(n-1) \wedge \mathbf{r}_\alpha(n-1)\}\\ &= \delta t\,\{\dot{\mathbf{r}}_\alpha(n-1) - \boldsymbol{\omega}(n-1) \wedge \mathbf{r}_\alpha(n-1)\},\end{aligned}$$

which becomes on a continuous time scale

$$\boldsymbol{\delta}\mathbf{u}_\alpha(t+\delta t) \approx \delta t\,\{\dot{\mathbf{r}}_\alpha(t) - \boldsymbol{\omega}(t) \wedge \mathbf{r}_\alpha(t)\}. \tag{11}$$

One can now construct the trajectory of the internal motions for a flexible macromolecule, by accumulating for each atom its virtual displacements in the course of time. Here, it must be taken into account that virtual displacements at different times t and $t' > t$ are related to virtual rigid molecules which have been defined at, $t - \delta t$ and $t' - \delta t$, respectively. Before adding these virtual displacements they should first be transformed to a common molecule-fixed frame, observing that the molecule has undergone a global rotation within the time interval $t' - t$. This point has not been considered in Ref. 10,

but it is important as will be shown below. The trajectory representing the internal motions of atom α then has the form

$$\mathbf{x}_\alpha^{int}(t) = \mathbf{x}_\alpha(0) + \int_0^t \mathbf{D}^T(\tau)\cdot\boldsymbol{\delta}\mathbf{u}_\alpha(\tau), \tag{12}$$

where $\mathbf{D}(t)$ is an orthogonal matrix describing the accumulated rotation of the macromolecule and "T" denotes a transposition. The matrix $\mathbf{D}(t)$ fulfills the differential equation

$$\dot{\mathbf{D}}(t) = \boldsymbol{\Omega}(t)\cdot\mathbf{D}(t), \tag{13}$$

where $\boldsymbol{\Omega}$ is the skew-symmetric matrix containing the Cartesian coordinates of angular velocity in the laboratory frame,

$$\boldsymbol{\Omega} = \begin{pmatrix} 0 & -\omega_z & \omega_y \\ \omega_z & 0 & -\omega_x \\ -\omega_y & \omega_x & 0 \end{pmatrix}. \tag{14}$$

Writing $\boldsymbol{\omega}\wedge\mathbf{r}_\alpha = \boldsymbol{\Omega}\cdot\mathbf{r}_\alpha$, one finds from the definition of the accumulated internal displacements that

$$\begin{aligned}\int_0^t \mathbf{D}^T(\tau)\cdot\boldsymbol{\delta}\mathbf{u}_\alpha(\tau) &= \int_0^t d\tau\,\mathbf{D}^T(\tau)\cdot\{\dot{\mathbf{r}}_\alpha(\tau) - \boldsymbol{\Omega}(\tau)\mathbf{r}_\alpha(\tau)\}\\ &= \left|\mathbf{D}^T(\tau)\cdot\mathbf{r}_\alpha(\tau)\right|_0^t - \int_0^t d\tau\,\left\{\underbrace{\dot{\mathbf{D}}^T(\tau)\cdot\mathbf{r}_\alpha(\tau) + \mathbf{D}^T(\tau)\cdot\boldsymbol{\Omega}(\tau)\mathbf{r}_\alpha(\tau)}_{=\mathbf{0}}\right\}\\ &= \mathbf{D}^T(t)\cdot\mathbf{r}_\alpha(t) - \mathbf{r}_\alpha(0).\end{aligned}$$

Here, it was used that $\dot{\mathbf{D}}^T = (\boldsymbol{\Omega}\cdot\mathbf{D})^T = \mathbf{D}^T\cdot\boldsymbol{\Omega}^T = -\mathbf{D}^T\cdot\boldsymbol{\Omega}$ and that $\mathbf{D}(0) = \mathbf{1}$. Inserting the above relation for the accumulated internal displacements into the definition (12) for the trajectory of internal motions leads to

$$\mathbf{x}_\alpha^{int}(t) = \mathbf{X}(0) + \mathbf{D}^T(t)\cdot\{\mathbf{x}_\alpha(t) - \mathbf{X}(t)\}, \tag{15}$$

which defines a roto-translation for the transformation $\mathbf{x}_\alpha(t)\rightarrow\mathbf{x}_\alpha^{int}(t)$. Such a transformation guarantees that the internal energy of the molecule is left invariant,

$$U(\mathbf{x}_1(t),\ldots,\mathbf{x}_N(t)) = U(\mathbf{x}_1^{int}(t),\ldots,\mathbf{x}_N^{int}(t)). \tag{16}$$

The matrix $\mathbf{D}(t)$ is the transformation matrix from the laboratory frame to a molecule-fixed frame, which can be considered as the Eckart frame of the molecule. Its elements define the projections of the moving molecule-fixed basis vectors $\boldsymbol{\epsilon}_i(t)$ $(i = 1, 2, 3)$ onto the fixed basis vectors $\mathbf{e}_j$ $(j = 1, 2, 3)$,

$$D_{ij}(t) = \boldsymbol{\epsilon}_i^T(t)\cdot\mathbf{e}_j. \tag{17}$$

While the center-of-mass vector $\mathbf{X}(t)$ can be easily obtained from a given molecular dynamics trajectory, the construction of the matrix $\mathbf{D}(t)$ is a less simple task that will be discussed in Sec. III.

III. CONSTRUCTING THE ROTATION MATRIX

This section describes the approximate construction of the rotation matrix $\mathbf{D}(t)$ and the corresponding trajectory of internal motions $\mathbf{x}_\alpha^{int}(t)$ from a molecular dynamics trajectory which is sampled with a finite time step Δt.

A. Discretization

In the general case, where $\boldsymbol{\Omega}$ is time dependent, only a formal solution of the defining Eq. (13) for $\mathbf{D}(t)$ can be given. It follows from Eq. (13) that

$$\mathbf{D}(t+\Delta t) = \mathbf{D}(t) + \int_t^{t+\Delta t} d\tau\,\boldsymbol{\Omega}(\tau)\cdot\mathbf{D}(\tau), \tag{18}$$

which may be approximated by

$$\mathbf{D}(t+\Delta t) \approx (\mathbf{1} + \Delta t\,\boldsymbol{\Omega}(t))\cdot\mathbf{D}(t), \tag{19}$$

if Δt tends to zero. Defining $\Delta t = t/n$, the rotation matrix is thus given by the infinite product

$$\mathbf{D}(t) = \lim_{n\rightarrow\infty}\prod_{k=0}^{n}\left(\mathbf{1} + \frac{t}{n}\,\boldsymbol{\Omega}([k/n]t)\right). \tag{20}$$

One recognizes that $\mathbf{D}(t) = \exp(\boldsymbol{\Omega}t)$ if $\boldsymbol{\Omega}$ is constant, i.e., if the molecule rotates with constant angular velocity.

For practical purposes, a discrete approximation of the matrix $\mathbf{D}(t)$ can be obtained from an approximation of expression (20),

$$\mathbf{D}(n) \approx \prod_{k=0}^{n-1}\boldsymbol{\Delta}\mathbf{D}(k), \tag{21}$$

where the increments $\boldsymbol{\Delta}\mathbf{D}(k)$ are given by

$$\boldsymbol{\Delta}\mathbf{D}(k) = \exp(\Delta t\,\boldsymbol{\Omega}(k)) \tag{22}$$

and appear in the *finite* roto-translation

$$\mathbf{x}_\alpha^{(c)}(k+1) = \mathbf{X}(k) + \boldsymbol{\delta}\boldsymbol{\tau}(k) + \boldsymbol{\Delta}\mathbf{D}(k)\cdot\mathbf{r}_\alpha(k), \tag{23}$$

which replaces relation (2) in the minimization problem (1). If $\mathbf{X}$ is chosen to be the center of mass, problem (1) takes the

form ($n = 0, 1, 2, \ldots$),

$$\xi = \frac{1}{2}\sum_{\alpha=1}^{N} m_\alpha \left(\mathbf{r}_\alpha(n+1) - \mathbf{r}_\alpha^{(c)}(n+1)\right)^2 = \text{Min}, \tag{24}$$

where the constrained positions with respect to the center of mass at time $n+1$ are obtained by a rotation from the positions at time n,

$$\mathbf{r}_\alpha^{(c)}(n+1) = \mathbf{\Delta D}(n) \cdot \mathbf{r}_\alpha(n). \tag{25}$$

Here, one may use that $\mathbf{r}_\alpha^{int}(n) = \mathbf{D}^T(n) \cdot \mathbf{r}_\alpha(n)$, which leads to the alternative expression

$$\mathbf{r}_\alpha^{(c)}(n+1) = \mathbf{D}(n+1) \cdot \mathbf{r}_\alpha^{int}(n), \tag{26}$$

observing that $\mathbf{D}(n+1) = \mathbf{\Delta D}(n) \cdot \mathbf{D}(n)$. Inserting either of the expressions (25) or (26) for $\mathbf{r}_\alpha^{(c)}(n+1)$ into the minimization problem (24) makes the target function ξ defined in Eq. (24) a function of the components of a rotation matrix, which must be chosen such that ξ is minimal.

B. Superposition problem

A convenient way to solve the rotational superposition problem (24) is to express the rotation matrix minimizing the target function ξ in terms of four real quaternion parameters. The general form for a rotation matrix in this parametrization is[14]

$$\mathbf{D}(q) = \begin{pmatrix} q_0^2 + q_1^2 - q_2^2 - q_3^2 & 2(-q_0q_3 + q_1q_2) & 2(q_0q_2 + q_1q_3) \\ 2(q_0q_3 + q_1q_2) & q_0^2 + q_2^2 - q_1^2 - q_3^2 & 2(-q_0q_1 + q_2q_3) \\ 2(-q_0q_2 + q_1q_3) & 2(q_0q_1 + q_2q_3) & q_0^2 + q_3^2 - q_1^2 - q_2^2 \end{pmatrix}, \tag{27}$$

where $q_0^2 + q_1^2 + q_2^2 + q_3^2 = 1$. Different variants for the solution of the rotational superposition problem with quaternions have been discovered and described by several authors (see, for example, Refs. 15–17 and the review Ref. 18) and more details may be found in Ref. 10. Using the method described in Ref. 17, the target function ξ given by relation[(24)] becomes a quadratic form in the quaternion parameters q,

$$\xi(q) = \frac{1}{2}\mathbf{q}^T \cdot \boldsymbol{\mu} \cdot \mathbf{q}, \tag{28}$$

If the coordinate sets to be superposed are $\{\mathbf{r}_\alpha(n)\}$ and $\{\mathbf{r}_\alpha(n+1)\}$ (see Eq. (25)), we have $q \equiv \Delta q(n)$ and furthermore define $\mathbf{r}_\alpha \equiv \mathbf{r}_\alpha(n)$. For the alternative form given by Eq. (26), we have $q \equiv q(n+1)$ and $\mathbf{r}_\alpha \equiv \mathbf{r}_\alpha^{int}(n)$ instead. For both cases we also define $\mathbf{r}'_\alpha \equiv \mathbf{r}_\alpha(n+1)$ and write the matrix $\boldsymbol{\mu}$ in the form

$$\boldsymbol{\mu} = \sum_{\alpha=1}^{N} m_\alpha \boldsymbol{\mu}_\alpha, \tag{29}$$

where the atomic contributions are given by

$$\boldsymbol{\mu}_\alpha = \begin{pmatrix} (\mathbf{r}_\alpha - \mathbf{r}'_\alpha)^2 & 2(\mathbf{r}'_\alpha \wedge \mathbf{r}_\alpha)^T \\ 2(\mathbf{r}'_\alpha \wedge \mathbf{r}_\alpha) & (\mathbf{r}_\alpha + \mathbf{r}'_\alpha)^2\mathbf{1} - 2(\mathbf{r}'_\alpha\mathbf{r}_\alpha^T + \mathbf{r}_\alpha\mathbf{r}'^T_\alpha) \end{pmatrix}. \tag{30}$$

Minimization of the target function $\xi(q)$ subject to the normalization constraint of the quaternion parameters leads to solving an eigenvector problem for a 4×4 matrix,

$$\boldsymbol{\mu} \cdot \mathbf{q} = \lambda\, \mathbf{q}. \tag{31}$$

Here, the eigenvalues represent the (mass-weighted) mean square superposition error and the normalized eigenvector corresponding to the smallest eigenvalue contains thus the quaternion parameters describing the optimal solution. In a more general context, the mass-weighting scheme of the atomic contributions to the total superposition error can be replaced by any set of positive numbers.

We note that the quaternion-based superposition method also yields a measure for the *orientational distance* of two molecular structures. If $\{\mathbf{r}_\alpha\}$ and $\{\mathbf{r}'_\alpha\}$ are the sets of atomic coordinates of the two structures under consideration, their orientational distance can be defined as their Euclidean distance divided by the corresponding maximum distance. The latter is given by the eigenvalue λ_{max} of the eigenvector problem (31) and using the definition of the matrix $\boldsymbol{\mu}$, the orientational distance is given by[19]

$$\Delta_\Omega = \sqrt{\frac{\mu_{11}}{\lambda_{max}}}. \tag{32}$$

By construction $0 \le \Delta_\Omega \le 1$.

C. Algorithmic considerations

If the choice (25) for $\mathbf{r}_\alpha^{(c)}(n+1)$ is used in the target function (24), the accumulated rotation matrix $\mathbf{D}(n)$ and the trajectory $\mathbf{r}_\alpha^{int}(n)$ are constructed as follows:

$$\begin{aligned} &\mathbf{r}_\alpha(n) \quad \mapsto \mathbf{r}_\alpha(n+1) \quad \text{yields} \quad \Delta q(n), \\ &\mathbf{D}(n+1) = \mathbf{D}(\Delta q(n))\mathbf{D}(n), \\ &\mathbf{r}_\alpha^{int}(n+1) = \mathbf{D}^T(n+1) \cdot \mathbf{r}_\alpha(n+1), \end{aligned} \tag{33}$$

setting $\mathbf{D}(0) = \mathbf{1}$. The explicit matrix multiplication in the accumulation of $\mathbf{D}$ and the corresponding accumulation of numerical errors can be avoided by minimizing the target

function (24) with the choice (26) for $\mathbf{r}_\alpha^{(c)}(n+1)$. This leads to

$$\begin{aligned} &\mathbf{r}_\alpha^{int}(n) \quad \mapsto \mathbf{r}_\alpha(n+1) \quad \text{yields} \quad q(n+1), \\ &\mathbf{D}(n+1) = \mathbf{D}(q(n+1)), \\ &\mathbf{r}_\alpha^{int}(n+1) = \mathbf{D}^T(n+1)\cdot\mathbf{r}_\alpha(n+1), \end{aligned} \tag{34}$$

with the starting point $\mathbf{r}_\alpha^{int}(0) = \mathbf{r}_\alpha(0)$.

Finally, we note that approximations for the angular velocity can be computed from the quaternion parameters Δq obtained from the superposition fits according to scheme (33). It follows from the expression (22) for the increment $\mathbf{\Delta D}(k)$ and the general form (27) for a rotation matrix expressed in quaternion parameters that (the time argument is dropped)

$$\begin{pmatrix} \Delta q_0 \\ \Delta q_1 \\ \Delta q_2 \\ \Delta q_3 \end{pmatrix} = \begin{pmatrix} \cos(\omega\Delta t/2) \\ \sin(\omega\Delta t/2)\, n_x \\ \sin(\omega\Delta t/2)\, n_y \\ \sin(\omega\Delta t/2)\, n_z \end{pmatrix}, \tag{35}$$

where ω is the modulus of $\boldsymbol{\omega}$ and n_x, n_y, n_z are the components of the unit vector $\mathbf{n} = \boldsymbol{\omega}/\omega$.

IV. APPLICATIONS

In the following, some examples will be discussed in which the method described above is compared with the usual approach of removing global motions from the molecular dynamics trajectories of macromolecules, which consists of aligning all configurations with a fixed reference structure. Two examples are considered:

(a) the dynamics of the TRP-cage (tryptophan-cage) miniprotein,[20] which consists of only 20 amino acids, and which is believed to be the smallest protein exhibiting a stable fold (code 2JOF of the protein data bank (PDB) Ref. 21),
(b) the folding of a polypeptide chain of the same sequence length ("C-tail," in the following), taken from the C-terminal region of the eukaryotic anti-association factor 6, and believed to be very mobile and mostly unstructured.

A. TRP-cage miniprotein

We performed one MD simulation of the TRP-cage molecule at ambient conditions and one at $T = 400\,\mathrm{K}$ and normal pressure, in order to enhance internal motions in the protein. As initial structure we used in both cases the crystallographic coordinates deposited in entry 2JOF of the PDB, adding the hydrogen atoms according to the known chemical bond structure in amino acids. The protein was immersed in a solvent of 7159 water molecules, choosing a cubic simulation box with a box length of 6 nm and periodic boundary conditions. All simulations were performed with the Molecular Modeling Toolkit,[22] using the AMBER99 force field,[23] which includes the TIP3P force field[24] for the simulation of water molecules. Coulomb interactions were treated with the method proposed by Wolf *et al.*,[25] using a cutoff radius of 1.4 nm. Tests of the method can be found in Refs. 26 and 27. For both temperatures, we performed equilibration runs of 100 ps at constant temperature and pressure in the *NpT* ensemble using the Nosé-Andersen extended system method,[28] each followed by a production run of 2 ns simulation at constant temperature in the *NVT* ensemble. For the integration of the equations of motion we used a time step of 1 fs and the protein configurations were stored every 20 fs for later analysis.

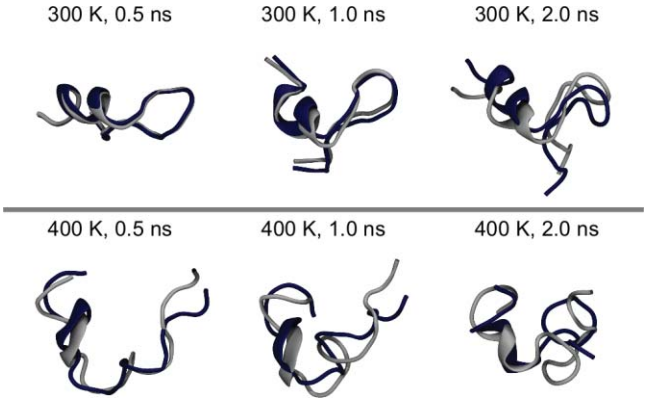

FIG. 1. Snapshots of the TRP-cage molecule taken from trajectories for the internal dynamics at 300 K and 400 K, which have been created by aligning the molecule (a) with the initial structure in the MD trajectory ("fit to first," snapshots in light gray) and (b) by employing the method presented in this paper ("fit to preceding," snapshots in dark blue).

Figure 1 displays some snapshots taken along the trajectories for internal motions which have been obtained by aligning the molecule with the first configuration in the production trajectory (light gray) and by using the method proposed in this paper (dark blue). In the following, the two approaches are referred to as "fit to first" and "fit to preceding," respectively. It should be noted that the resulting coordinate sets differ by a global rigid-body motion, if the same time frames are compared, and the internal energy is thus strictly the same. The deviation of the global orientation increases, however, rapidly with time, and this effect is more pronounced for the strongly heated protein at $T = 400\,\mathrm{K}$. To quantify this deviation we display in Fig. 2 the corresponding orientational distance which is defined by relation (32). More important for applications are the resulting differences for time averaged quantities, such as static averages, time-correlation functions and time-dependent mean square displacements (MSD). The first comparison of the different methods to remove global

TABLE I. Total (mass-weighted) RMSF and $\lim_{t\to\infty} W(t)$ values computed for the TRP-cage molecule (at 300 and 400 K) and for the C-tail peptide. In the first column, "ff" stands for the "fit to first" method and "fp" stands for the "fit to preceding" method proposed in this paper.

Simulation	Δ [nm]	$\lim_{t\to\infty} W(t)$ [nm^2]
TRP 300 K ff	0.239	0.115
TRP 300 K fp	0.241	0.116
TRP 400 K ff	0.407	0.331
TRP 400 K fp	0.432	0.372
C-tail 300 K ff	0.553	0.612
C-tail 300 K fp	0.581	0.676

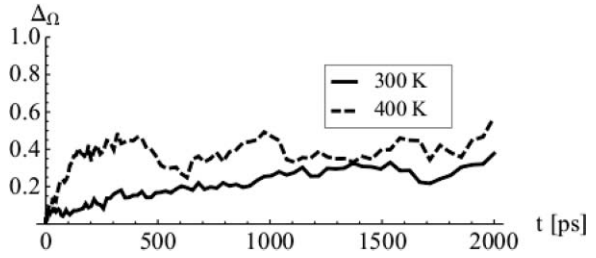

FIG. 2. Time evolution of the orientational distance between the molecular configurations of the TRP-cage molecule generated by the methods "fit to first" and "fit to preceding." The solid line corresponds to $T = 300$ K and the dashed line to $T = 400$ K.

motions concerns the root mean square fluctuations (RMSF) of the atoms in the TRP-cage protein. We define the total RMSF as

$$\Delta = \sqrt{\sum_{\alpha=1}^{N} w_\alpha \langle (\mathbf{r}_\alpha - \langle \mathbf{r}_\alpha \rangle)^2 \rangle}, \tag{36}$$

where $\langle \ldots \rangle$ denotes an average over the MD trajectory and w_α are atomic weights with $\sum_{\alpha=1}^{N} w_\alpha = 1$ (N is the number of atoms in the protein). In our calculations we use mass-weighting, i.e., $w_\alpha = m_\alpha / M$, with M being the total mass of the protein. The RMSF values are displayed in the second column of Table I. The RMSF values are almost equal at $T = 300$ K, but at higher temperature ($T = 400$ K) the difference is much more pronounced. So, we note that at $T = 400$ K, the "fit to preceding" method leads to a greater flexibility of the protein compared to the "fit to first" method.

The second comparison of the different methods concerns the time-dependent mean-square displacement which is defined as

$$W(t) = \sum_{\alpha=1}^{N} w_\alpha \langle [\mathbf{r}_\alpha(t) - \mathbf{r}_\alpha(0)]^2 \rangle, \tag{37}$$

where w_α are atomic weights with the same definition as mentioned previously. It measures how far (on average) the system moves away from its original configuration in a given amount of time and for confined diffusion the plateau value is related to the RMSF,

$$\lim_{t \to \infty} W(t) = 2\,\Delta^2. \tag{38}$$

Figure 3 displays the mass-weighted average MSDs for the TRP-cage molecule corresponding, respectively, to the "fit to first" and "fit to preceding" trajectory. The comparison shows that practically identical results are obtained for $T = 300$ K, whereas differences appear for 400 K. Although these differences are small in amplitude, they are not irrelevant since the form of the MSD changes. The asymptotic values $\lim_{t\to\infty} W(t)$ are shown in the third column of Table I next to the values of Δ from which they were computed. As we could expect from the trends of the MSD curves during the first picoseconds (Fig. 3), the $W(\infty)$ values are similar at $T = 300$ K and a little higher for the "fit to preceding" method than for the "fit to first" method at 400 K.

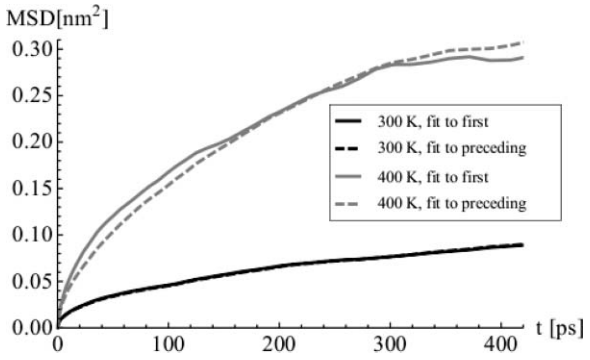

FIG. 3. Average (mass-weighted) MSD for the internal motions of the TRP-cage molecule for $T = 300$ K (black lines) and $T = 400$ K (gray lines). For both temperatures the MSD is calculated from trajectories created by the methods "fit to first" and "fit to preceding" (solid line and dashed line, respectively).

B. C-tail

The C-tail polypeptide chain (sequence EDAQPESISGNLRDTLIETYS in the one-letter code notation for residue types) was simulated starting from a linear configuration created by using the AMBER9 (Ref. 29) simulation package. All equilibration and production runs were performed using the AMBER03 (Ref. 30) force field and an implicit solvent. The latter is described by the Generalized Born solvation model, developed by Hawkins *et al.*,[31,32] where mean forces are obtained from the estimation of the total solvation free energy of the molecule into water. Initial equilibration of the linear structure was performed with progressive temperature rescaling from 0 K to 300 K with an increase of 50 K every 500 ps. The time step during this initial equilibration was varied from 0.1 to 0.5 fs, in order to reduce the extent of force variation and thus the probability of unphysical close contacts. After a final equilibration simulation of 700 ps, performed at constant temperature (300 K) and constant volume, the production run of 40 ns was performed with an integration time-step of 1 fs in the *NVT* ensemble. Configurations of the C-tail polypeptide chain were saved every 500 fs for later analysis.

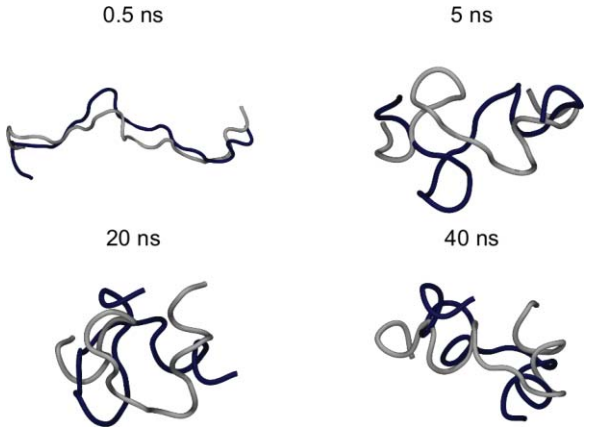

FIG. 4. Snapshots of the C-tail peptide. The coloring scheme is the same as in Fig. 1.

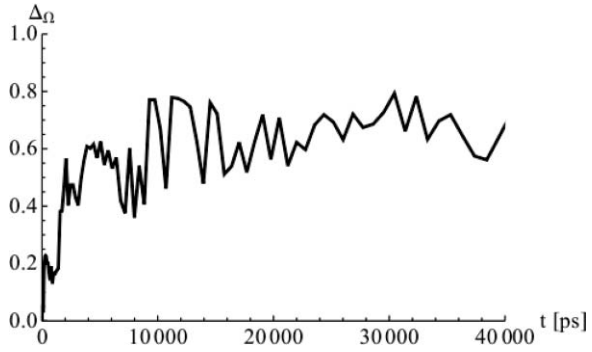

FIG. 5. Time evolution of the orientational distance between the molecular configurations of the C-tail peptide generated by the methods "fit to first" and "fit to preceding."

Some snapshots of the folding path of C-tail are displayed in Fig. 4, where the configurations obtained by aligning the polypeptide with the "fit to first" method (light gray) and by using the "fit to preceding" method (dark blue) are shown at 0.5, 5, 20, and 40 ns. The progressively divergent configurations obtained by these two methods become well distinguishable after a few nanoseconds as also suggested by the orientational distance Δ_Ω between corresponding configurations obtained by the two methods (see Fig. 5). This divergence is due to the fact that the reference structure for the "fit to first" method is almost completely extended and thus very different from the folded structures in the trajectory. Here, the reader should note that the slowly increasing value of Δ_Ω from zero to a plateau value of 0.6 is strictly related to the folding evolution: as the polypeptide reaches a more or less folded conformation, Δ_Ω starts oscillating around an average value. The latter fact is due to the negligible difference between the two methods when applied to folded conformations.

As for the TRP-cage simulation, we calculated the mass-weighted average RMSDs (Table I) and MSDs (Fig. 6) in order to explore the difference of the two fitting methods on the internal dynamics of C-tail. As for the TRP-cage at 400 K, we note that the method proposed in this work leads to a larger fluctuations compared to the conventional "fit to first" method, the difference being even more striking, as is to be expected given the large orientational differences between the two trajectories.

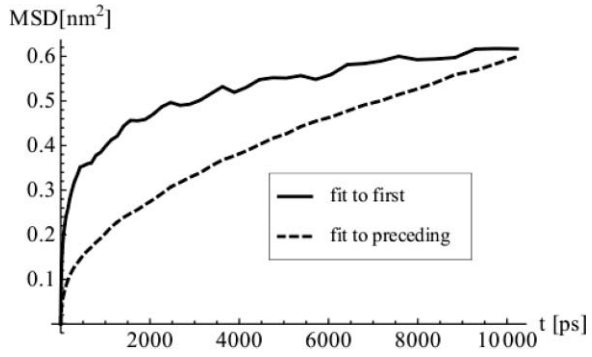

FIG. 6. Average (mass-weighted) atomic mean square displacement for the internal motions of the C-tail peptide computed from trajectories generated by the methods "fit to first" and "fit to preceding."

V. CONCLUSIONS

It has been shown that the construction of internal trajectories for flexible macromolecules from a given molecular dynamics trajectory can be achieved in a systematic way by accumulating the virtual internal displacements obtained from Gauss' principle of least constraint in a molecule-fixed frame. For a given time and a short time increment, these virtual displacements are defined as the differences between the actual displacements of the atoms in the molecule and those obtained from a corresponding rigid-body motion within the same time interval. The translation of the molecule-fixed frame is determined by the translational motion of the center of mass of the molecule and its rotation is described by the accumulated infinitesimal rotations of the virtual rigid bodies which are used to define the internal atomic displacements. The coordinate transformation describing the extraction of internal motions from the input trajectory can be written as a roto-translation of the molecule and thus strictly preserves its internal potential energy.

The application to the dynamics of TRP-cage miniprotein around its equilibrium conformation at ambient conditions shows that the standard procedure for removing global motions from the trajectory of a macromolecule, which consists of aligning its configurations with a fixed reference structure, yields results that are practically identical with the procedure proposed in this paper. When the protein is strongly heated, such that it starts to explore non-native conformations, the analysis of the respective trajectories displays slight differences. If a folding process is considered, where the conformation of the molecule under consideration changes considerably during the simulation, the comparison with a fixed reference structure is not appropriate for extracting internal motions and the method proposed in this paper should be used instead. In this method, least-square fit alignments are always made between neighboring steps in the trajectory, and thus between configurations that differ only slightly. The method does not depend on the arbitrary choice of a reference configuration and yields the exact solution in the limit $\Delta t \to 0$. Since the computational effort is the same as for the standard method, our method can replace it safely and efficiently in all situations to extract the internal motions of macromolecules from a given molecular dynamics trajectory.

[1] R. Abseher, S. Lüdemann, H. Schrieber, and O. Steinhauser, J. Mol. Biol. **249**, 604 (1995).
[2] J. Chen, C. Brooks, and P. Wright, J. Biomol. NMR **29**, 243 (2004).
[3] A. Nederveen and A. Bonvin, J. Chem. Theory Comput. **1**, 363 (2005).
[4] V. Calandrini, D. Abergel, and G. R. Kneller, J. Chem. Phys. **133**, 145101 (2010).
[5] G. Kneller and V. Calandrini, Biochim. Biophys. Acta **1804**, 56 (2010).
[6] C. Eckart, Phys. Rev. **47**, 552 (1935).
[7] A. Sayvetz, J. Chem. Phys. **7**, 383 (1939).
[8] K. Kudin and A. Dymarsky, J. Chem. Phys. **122**, 224105 (2005).
[9] C. Gauss, J. Reine Angew. Math. **4**, 232 (1829).
[10] G. Kneller, J. Chem. Phys. **128**, 194101 (2008).
[11] L. Pars, *A Treatise on Analytical Dynamics* (Heinemann, London, 1965).

[12]C. Lanczos, *The variational Principles of Mechanics*, 4th ed. (University of Toronto, Toronto, 1974).
[13]G. Kneller, J. Chem. Phys. **127**, 164114 (2007).
[14]S. Altmann, *Rotations, Quaternions, and Double Groups* (Clarendon, Oxford, 1986).
[15]R. Diamond, Acta Crystallogr. A **44**, 211 (1988).
[16]R. Diamond, Acta Crystallogr. A **46**, 423 (1990).
[17]G. Kneller, Mol. Simul. **7**, 113 (1991).
[18]D. Flower, J. Mol. Graph. Model. **17**, 238 (1999).
[19]G. Kneller and P. Calligari, Acta Crystallogr. D **62**, 302 (2006).
[20]B. Barua, J. C. Lin, V. D. Williams, P. Kummler, J. W. Neidigh, and N. H. Andersen, Protein Eng. Des. Sel. **21**, 171 (2008).
[21]J. Kirchmair, P. Markt, S. Distinto, D. Schuster, G. M. Spitzer, K. R. Liedl, T. Langer, and G. Wolber, J. Med. Chem. **51**, 7021 (2008).
[22]K. Hinsen, J. Comput. Chem. **21**, 79 (2000); see URL http://dirac.cnrs-orleans.fr.
[23]J. Wang, P. Cieplak, and P. Kollman, J. Comput. Chem. **21**, 1049 (2000).
[24]W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein, J. Chem. Phys. **79**, 926 (1983).
[25]D. Wolf, P. Keblinski, S. Philpot, and J. Eggebrecht, J. Chem. Phys. **110**, 8254 (1999).
[26]A. Angoshtari and A. Yavari, Phys. Lett. A **375**, 1281 (2011).
[27]D. A. C. Beck, R. S. Armen, and V. Daggett, Biochemistry **44**, 609 (2005).
[28]S. Nosé, J. Chem. Phys. **81**, 511 (1984).
[29]D. Case, T. Cheatham III, T. Darden, H. Gohlke, R. Luo, K. M. Merz, Jr., A. Onufriev, C. Simmerling, B. Wang, and R. Woods, J. Comput. Chem. **26**, 1668 (2005).
[30]Y. Duan, C. Wu, S. Chowdhury, M. Lee, G. Xiong, W. Zhang, R. Yang, P. Cieplak, R. Luo, and T. Lee, J. Comput. Chem. **24**, 1999 (2003).
[31]G. Hawkins, C. Cramer, and D. Truhlar, Chem. Phys. Lett. **246**, 122 (1995).
[32]G. Hawkins, C. Cramer, and D. Truhlar, J. Phys. Chem. **100**, 19824 (1996).