JCTC Journal of Chemical Theory and Computation

Article

pubs.acs.org/JCTC

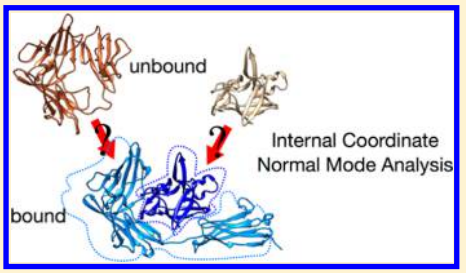
Internal Normal Mode Analysis (iNMA) Applied to Protein Conformational Flexibility

Elisa Frezza and Richard Lavery*

BMSSI, UMR 5086 CNRS/Univ. Lyon I, Institut de Biologie et Chimie des Protéines, 7 passage du Vercors, Lyon 69367, France

Supporting Information

ABSTRACT: We analyze the capacity of normal modes to predict observed protein conformational changes, and, notably, those induced by the formation of protein−protein complexes. We show that normal modes calculated in internal coordinate space (ICS) provide better predictions. For a large test set, using the ICS approach describes the conformational changes more completely, and with fewer low-frequency modes than the equivalent Cartesian coordinate modes, despite the fact that the internal coordinate calculations were restricted to torsional angles. This can be attributed to the fact that the use of ICS extends the range over which movements along the corresponding eigenvectors remain close to the true conformational energy hypersurface. We also show that the PaLaCe coarse-grain protein model performs better than a simple elastic network model. We apply ICS normal-mode analysis to protein complexes and, by extending the approach of Sunada and Gō, [Sunada, S.; Gō, N. *J. Comput. Chem.* **1995**, 16, 328−336], we show that we can couple an accurate view of the Cartesian coordinate movements induced by ICS modes with the detection of the key residues responsible for the movements.



1. INTRODUCTION

Knowledge of the structure of protein−protein complexes and their interactions is of utmost importance for understanding molecular mechanisms in biology.[1−4] However, only a small fraction of protein−protein complexes have been experimentally determined so far.[5,6] Hence, there is considerable interest in using computational techniques to help fill this gap.[7] Protein docking algorithms aim at predicting the three-dimensional structures of unknown complexes starting from the structures of the individual subunits.[8,9] However, proteins are flexible and they can undergo substantial conformational rearrangements upon binding[10] that are thus crucial for their function.[11] The major challenge of docking approaches is thus identifying correct solutions, while properly dealing with molecular flexibility.

In principle, a powerful and computationally inexpensive method for dealing with protein flexibility is offered by normal mode calculations.[12−16] Normal mode analysis (NMA) provides information on the equilibrium modes accessible to a system, within a harmonic approximation. Following many decades of applications to classical physical problems, molecular applications confirmed the importance of low-frequency motions in biological processes. There two main strategies for computing normal modes, depending on the description of the degrees of freedom: Cartesian coordinate space (CCS) and internal coordinate space (ICS). The former is computationally simpler and is the most popular approach. It is typically used in conjunction with coarse-grain (CG) protein models that often represent only a simplified protein backbone (i.e., the Cartesian positions of the Cα atoms) with Hook's law springs joining the pairs of Cα atoms closer than a chosen cutoff distance. This choice is exemplified by the usual implementation of the Anisotropic Network Model (ANM) approach,[17−19] Other approaches also use CCS normal modes, but include more structural information on both the backbone and the side-chain atoms (see, for example, refs 20−22).

Although NMA has been widely used to understand allosteric processes,[23,24] and also protein deformations induced by ligand binding,[19,25−27] there have been few studies using normal modes to describe the conformational changes between unbound (that is, isolated) and bound proteins.[28−33] Zacharias and collaborators investigated the use of normal modes obtained from ANM calculations on unbound proteins to account for conformational changes during protein−protein docking, but concluded that it was not clear if inclusion of the soft, collective ANM modes could improve the performance of systematic docking simulations, without prior knowledge of the binding geometry.[30,32] Dobbins et al. applied CCS NMA to a set of 20 proteins and compared low-frequency modes with the unbound-to-bound transitions.[29] They showed that a single low-frequency normal mode could satisfactorily describe the observed conformational changes in roughly only 1/3 of the cases under investigation. Stein et al. considered a larger benchmark of 2090 unbound-to-bound transitions using an elastic network CCS NMA considering only Cα atoms. They also found that ∼1/3 of the proteins studied were likely to explore the bound conformation without the presence of any interacting partner, thus strongly supporting the conformational selection model.[33]

What are the problems with normal modes? Most importantly, they model thermally induced conformational fluctuations with

Received: July 30, 2015
Published: September 30, 2015

DOI: 10.1021/acs.jctc.5b00724
J. Chem. Theory Comput. 2015, 11, 5503−5512

a harmonic approximation to the energy hypersurface. In contrast, protein deformations induced by ligand or protein binding may be larger, fueled by energy available during the formation of the interactions.[12,34−37] However, it has been pointed out that the motions described by low-frequency eigenvectors are more useful than data derived from the eigenvalues for interpreting the functionally relevant motions.[38,39] The results cited above show that improvements must be made if NMA is to become a useful predictive tool for binding-induced protein deformations.

One clear route for improving NMA involves the choice of the coordinate space. More than two decades ago, Gō and collaborators noted that ICS is more advantageous, since it extends the validity of the assumed harmonicity of conformational energy hypersurface.[40] Thus, the eigenvectors of low-frequency modes can be used to model larger conformational changes. In addition, the choice "chemically relevant" variables (torsion angles, valence angles, and bond lengths) makes it easy to partition degrees of freedom into "hard" and "soft" and to optionally simplify and accelerate computations by excluding variables that are energetically unlikely to make major contributions to large, collective movements. Thus, is it easy to perform ICS NMA in torsional angle space (TAS) by ignoring valence angle and bond length variations. It is also equally possible to perform the analysis on any relevant subset of variables without changing the molecular representation of the system under study.

Despite these advantages, ICS NMA has been rarely used, possibly because it is more complex to program and it is not available in the common molecular modeling packages. The complexity comes from the fact that, in ICS, internal variables must not be mixed with overall translations or rotations of the system under study. This requires information on the topology of the molecular system studied and notably on which particles are moved by any given variable. While, in single molecules, these problems can be restricted to describing torsion angle moves, molecular complexes require the treatment of both valence angles and bond lengths to describe interbody degrees of freedom.[41−43] Lastly, while ICS normal modes provide very useful information on which variables are responsible for low-frequency collective movements, they do not spontaneously describe the movements occurring in Cartesian space. Therefore, having a CCS description while conserving the advantages of the ICS modes is certainly interesting.[44−46] This problem has been solved for the case of the single protein up to a second-order expansion,[44,47] but further work was necessary to be able to treat protein complexes (or molecular assemblies in general), and this has been done as part of the current study.

Here, we will use NMA to study unbound-to-bound transitions in proteins, and also the dynamics of protein complexes, in conjunction with the PaLaCe CG protein model.[48] However, we should stress that our main objective is to compare the applicability of CCS and ICS normal mode approaches and not to test the qualities of the PaLaCe CG representation or its associated force field. The code we have developed and the methodological extensions we have made are perfectly applicable to any molecular representation and to any force field (provided an accurate energy minimum can be generated and both first and second derivatives of the energy, with respect to the conformational degrees of freedom can be obtained). However, we do make a limited comparison of PaLaCe CCS results with those obtained using a simple elastic network, since this model has been very popular over recent years and has already been used for analyzing unbound-to-bound transitions.

The manuscript is structured as follows. The next section outlines NMA, emphasizing the major features of the internal coordinate approach and derives the equations necessary for the conversion from ICS to CCS. Then, we discuss computational details with particular reference to the CG model that we use and the protocol for calculating normal modes and the conformational changes between unbound and bound proteins. We then contrast CCS and ICS calculations applied to unbound-to-bound conformational transition proteins for a large and varied set of proteins. Finally, we draw the conclusions of this study and discuss the applicability of the ICS approach.

2. NORMAL MODE ANALYSIS

2.1. General Theory. Normal mode theory[49] provides an analytical solution of the equations of motion by imposing a harmonic approximation on the potential energy of the system (i.e., by assuming that the energy is a quadratic function of its $\mathcal{N}$ coordinates) around the potential energy minimum q_i^0:

$$E_p = \frac{1}{2}\mathbf{q}^{\mathrm{T}}\mathbf{F}\mathbf{q} \tag{1}$$

where $\mathbf{q}$ is a set of coordinates, $\mathbf{F}$ is the potential energy matrix, or Hessian matrix, defined by $F_{ij} = \partial^2 E_p/(\partial q_i \partial q_j)$. The kinetic energy of the molecule can be expressed in terms of internal variables:

$$E_k = \frac{1}{2}\dot{\mathbf{q}}^{\mathrm{T}}\mathbf{H}\dot{\mathbf{q}} = \frac{1}{2}\sum_a m_a \dot{\mathbf{q}}^{\mathrm{T}} \frac{\partial \mathbf{r}_a}{\partial q_i}\frac{\partial \mathbf{r}_a}{\partial q_j}\dot{\mathbf{q}} \tag{2}$$

where $H_{ij} = \sum_a m_a(\partial \mathbf{r}_a/\partial q_i)(\partial \mathbf{r}_a/\partial q_j)$ is the kinetic matrix, m_a represents the atomic masses, and $\mathbf{r}_a$ represent the position vectors of each atom a; we can also define the matrix $\mathbf{K}$ with $K_{ij} = \sum_a(\partial \mathbf{r}_a/\partial q_i)(\partial \mathbf{r}_a/\partial q_j)$. The internal coordinates $\mathbf{q}$ can be rigid-body translations and rotations, torsional or valence angles and bond lengths. The equations of motion in terms of any set of coordinates are given by Lagrange's equations, whose solution takes the following form:

$$q_i(t) = q_i^0 + \sum_{k=1}^{n} A_{ik}Q_k = q_i^0 + \sum_{k=1}^{n} A_{ik}\alpha_k \cos(\omega_k t + \delta_k) \tag{3}$$

where α_k and δ_k are dependent on the initial conditions and are the amplitude and the phase of the kth mode, respectively. The unknowns, A_{ik} and ω_k, are obtained by solving the generalized eigenvector problem:

$$\mathbf{HAW} = \mathbf{FA} \tag{4}$$

where $\mathbf{W}$ is a diagonal positive matrix whose elements are

$$W_{ik} = \delta_{ik} w_k^{\,2} \tag{5}$$

and δ_{ik} is the Kronecker delta.

2.1.1. Thermal Amplitude. The amplitude α_k of a particular normal mode k is dependent on the temperature. Based on the equipartition theorem, each normal mode has a time-averaged potential energy equal to $^1/_2 k_B T$, where k_B is the Boltzmann constant and T is the absolute temperature. The time-average potential energy of each mode can be written as

$$\langle E_p \rangle = \frac{1}{2}\omega_k^2 \langle [\alpha_k \cos(\omega_k t + \delta_k)]^2 \rangle = \frac{1}{4}\omega_k^{\,2}\alpha_k^{\,2} \tag{6}$$

Hence, we obtain

$$\alpha_k = \frac{\sqrt{2k_B T}}{\omega_k} \tag{7}$$

DOI: 10.1021/acs.jctc.5b00724
J. Chem. Theory Comput. 2015, 11, 5503−5512

2.1.2. Time-Averaged Properties. Time-averaged properties of motion are dependent only on the amplitude α_k. The correlation coefficient is given by $\langle \delta q_i(\tau)\delta q_j(\tau)\rangle$:

$$\langle \delta q_i(\tau)\delta q_j(\tau)\rangle = \langle \sum_k A_{ik}(\tau) \sum_l A_{jl}Q_l(\tau)\rangle = \sum_k \sum_l A_{ik}A_{jl}\delta_{kl}\langle Q_k(\tau)Q_l(\tau)\rangle \quad (8)$$

Therefore, if $k = l$, $\langle Q_k(\tau)Q_l(\tau)\rangle = {\alpha_k}^2/2$. Thus,

$$\langle \delta q_i(\tau)\delta q_j(\tau)\rangle = \frac{1}{2}\sum_k A_{ik}A_{jk}{\alpha_k}^2 \quad (9)$$

2.1.3. Units and Conversion Factors. In these calculations, energy and second derivatives of potential energy are expressed in units of kcal mol^{-1}. Masses and the atomic displacements are expressed in units of g mol^{-1} and Å, respectively. Therefore, the elements H_{ij} of the kinetic matrix are expressed in units of g mol^{-1} Å^2, or in s^2 kcal mol^{-1} (4.1855×10^{26}). Consequently, the frequencies ω_k are expressed in $(4.1855)^{1/2} \times 10^{13}$s^{-1} (with this choice, 1 unit time is equal to $t_0 = 4.8889 \times 10^{-14}$ s, or ~49 fs). We can also convert the frequencies to cm^{-1} by multiplying them by $(4.1855)^{1/2} \times 10^{13}k/(2\pi c) = 108.6117$, where c is the speed of light (2.9979×10^{10} cm s^{-1}). We can also determine the oscillation period expressed in picoseconds, using the following expression:

$$T\ (\text{ps}) = \frac{10^{12}}{c\omega_k\ (\text{cm}^{-1})} \quad (10)$$

Finally, for the calculation of the thermal amplitude, we used eq 7 and we expressed k_B in units of kcal mol^{-1} K^{-1}, the temperature being given in Kelvin (K) and the frequency given in units of time.

2.2. Internal Coordinate Normal Analysis. *2.2.1. Internal Coordinates.* In the presence of N proteins, we must consider two types of variables: interbody (see Figure 1) and intrabody coordinates (here, limited to torsions and valence angles). The number of independent interbody variables is $6N - 6$. For the moment, we consider the atoms belonging to protein i to be a rigid body. For each body, we chose the atom closest to the center of mass as the pivot for the rotation and translation. To be able to study more than two proteins, we take one body as a reference system and keep it fixed in space. We express the resulting $N - 1$ interbody degrees of freedom as one relative translation (e.g., t_{12}) and one relative rotation (e.g., τ_{12}), plus one torsion and one bend angle (e.g., τ_2 and θ_1, respectively) per mobile protein, and $N - 1$ torsions and $N - 1$ bend angles for the static reference protein.

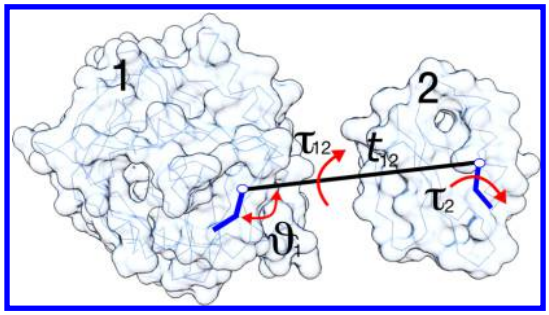


Figure 1. Definition of interbody variables for two rigid-body proteins labeled 1 and 2. Legend: t_{12}, relative translation; τ_{12}, relative rotation; θ_1 and θ_2, bending angles of proteins 1 and 2, respectively; τ_1 and τ_2: torsion angles of proteins 1 and 2, respectively.

2.2.2. Hessian Matrix. To obtain the Hessian matrix, we calculate the first derivative of the energy analytically and the second derivatives numerically (for computational efficiency). The first derivatives of the potential energy, with respect to the variables (torsions, valence angles, and bond lengths), can be written as

$$\frac{dE_p}{dq_i} = \sum_k \frac{dE_k}{d\mathbf{r}_k}\cdot\frac{d\mathbf{r}_k}{dq_i} \quad (11)$$

Here, in the case of torsions $d\mathbf{r}_j/d\tau_i = (\mathbf{b}_i \times \mathbf{R}_{ji})$, where $\mathbf{R}_{ji} = \mathbf{r}_j - \mathbf{r}_i$ and $\mathbf{b}_i$ is the unit bond vector $(\mathbf{b}_i = (\mathbf{r}_{i+1} - \mathbf{r}_i)/s_i$, with $s_i = |\mathbf{r}_{i+1} - \mathbf{r}_i|)$; in the case of the valence angles, $d\mathbf{r}_j/d\theta_i = (\mathbf{R}_{ji} \times \mathbf{c}_i)$, where $\mathbf{c}_i = (\mathbf{b}_{i-1} - \mathbf{b}_i)/\sin\theta_i$ is the perpendicular vector to the plane of the valence angle; in the case of the distances, $d\mathbf{r}_j/dt_i = \mathbf{b}_i$.

2.2.3. Kinetic Matrix. Concerning the kinetic matrix, $\mathbf{H}$, special attention must be paid to the derivatives of the atomic position vector $\mathbf{r}_a$, with respect to the independent variables q_i. Since the internal kinetic energy cannot include external motions (i.e., the overall translation and rotation of the system), the change $\delta\,\mathbf{r}_a$ caused by δq_i must not move the center of mass or modify the inertia tensor. To achieve this, Noguti and Gō proposed an elegant analytical calculation of $\mathbf{H}$ in torsion angle space (TAS).[41,42,47] For a given torsion angle, taking all the other variables to be fixed, the system can be treated as two rigid bodies connected by a chemical bond around which a rotation τ_p can occur. For each atom b in the first body and each atom c in the second, $\delta\,\mathbf{r}_b$ and $\delta\,\mathbf{r}_c$ are expressed using the Eckart conditions[50] to separate internal and external motion. These conditions can be formulated as

$$\sum_a m_a\Delta\mathbf{r}_a = 0 \qquad \sum_a m_a\mathbf{r}_a^0 \times \Delta\mathbf{r}_a = 0 \quad (12)$$

where m_a is the mass of atom a and $\mathbf{r}_a^0$ is the fixed position vector of atom a in the reference conformation of the molecule(s), usually corresponding to a minimum-energy conformation. We denote the total mass of the system by M, the mass and the center of mass of the ith rigid body as M_{ip} and $\mathbf{G}_{ip}^0$, respectively, and the total inertia tensor as $\mathbf{I}$. All these quantities are defined using the following expressions:

$$M_{1p} = \sum_{b\in M_\beta} m_b, \quad M_{2p} = \sum_{c\in M_\beta} m_c, \quad M = M_{1p} + M_{2p}$$

$$\mathbf{G}_{1p}^0 = \sum_{b\in M_\beta} \frac{m_b\mathbf{r}_b^0}{M_{1p}} \quad \mathbf{G}_{2p}^0 = \sum_{c\in M_\gamma} \frac{m_c\mathbf{r}_c^0}{M_{2p}}$$

$$I_{ij,1p} = \sum_{b\in M_\beta} m_b\left\{\delta_{ij}\sum_{k=1}^{3}(\mathbf{r}_{kb}^0)^2 - \mathbf{r}_{ib}^0\mathbf{r}_{jb}^0\right\}$$

$$I_{ij,2p} = \sum_{c\in M_\gamma} m_c\left\{\delta_{ij}\sum_{k=1}^{3}(\mathbf{r}_{kc}^0)^2 - \mathbf{r}_{ic}^0\mathbf{r}_{jc}^0\right\}$$

$$\mathbf{I} = \mathbf{I}_{1p} + \mathbf{I}_{2p} \quad (13)$$

The effect of varying the torsional angle τ_p by $\Delta\tau_p$ can be considered in two steps. First, rigid body 1 is fixed in space and only rigid body 2 is rotated around the bond p along the unit

DOI: 10.1021/acs.jctc.5b00724
J. Chem. Theory Comput. 2015, 11, 5503−5512

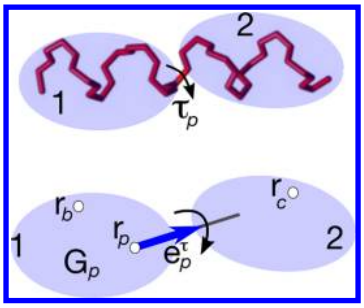


Figure 2. (Top) A molecule considered as two bodies 1 and 2 connected by a chemical bond p described and modified by the rotation of a torsional angle τ_p. (Bottom) $\mathbf{G}_p$ and $\mathbf{r}_p$ are the position vectors of the center of mass of rigid body 1 and of the atom at an end of bond p in rigid body 1, respectively. $\mathbf{e}^{\tau}_p$ is the unit vector in the direction of bond p. The position vectors of the atoms belonging to rigid bodies 1 and 2 are denoted by $\mathbf{r}_b$ and $\mathbf{r}_c$, respectively.

vector $\mathbf{e}^{\tau}_p$ by an infinitesimal amount $\Delta\tau_p$ (see Figure 2). The unit vector $\mathbf{e}^{\tau}_p$ is the axis of rotation for the torsional angle τ_p. In the following, we will denote the name of the bond involved by a subscript and the type of variable we considered by a superscript, where t represents a translation and τ or ψ represents a rotation. All the atoms belonging to rigid body 2 ($\mathbf{r}_c$) move by

$$\Delta\mathbf{r}^*_c = \mathbf{e}^{\tau}_p \times (\mathbf{r}^0_c - \mathbf{r}^0_p)\Delta\tau_p + \frac{1}{2}\mathbf{e}^{\tau}_p \times [\mathbf{e}^{\tau}_p \times (\mathbf{r}^0_c - \mathbf{r}^0_p)]\Delta\tau_p\Delta\tau_p \quad (14)$$

Then, the entire system is rotated by the rotation vector $\mathbf{\Omega}$, which passes through the center of mass $\mathbf{G}^0_{ip}$ and is translated by $\Delta\mathbf{Y}_1$ (in the following, α and β denote the first and second order in $\mathbf{\Omega}$, and $\Delta\mathbf{Y}^1$ and $\Delta\mathbf{Y}^2$ the first and second order in $\Delta\mathbf{Y}_1$). For this motion, we can write

$$\Delta\mathbf{r}_b = \Delta\mathbf{Y}_1 + \mathbf{\Omega} \times (\mathbf{r}^0_b - \mathbf{G}^0_{ip}) + \frac{1}{2}\mathbf{\Omega} \times [\mathbf{\Omega} \times (\mathbf{r}^0_b - \mathbf{G}^0_{ip})]$$

$$\Delta\mathbf{r}_c = \Delta\mathbf{Y}_1 + \mathbf{\Omega} \times (\mathbf{r}^0_c + \Delta\mathbf{r}^*_c - \mathbf{G}^0_{ip}) + \frac{1}{2}\mathbf{\Omega} \times [\mathbf{\Omega} \times (\mathbf{r}^0_b + \Delta\mathbf{r}^*_c - \mathbf{G}^0_{ip})] + \Delta\mathbf{r}^*_c \quad (15)$$

Next, we can determine the unknown quantities $\mathbf{\Omega}$ and $\Delta\mathbf{Y}_1$ satisfying the Eckart conditions, eqs 12. This approach in two steps can be generalized for N proteins considering both intra and inter rotations and distances.[41] It is important to note that, in the case of N proteins, a rigid body can be a fragment of a protein or an entire protein.

Inter and Intra Distances. We consider the case where a translation (or the modification of one bond length) is involved. We translate rigid body 2 along the unit vector $\mathbf{e}^t_p$ by an infinitesimal quantity Δt_p, all the atoms belonging to rigid body 2 ($\mathbf{r}_c$) move by

$$\Delta\mathbf{r}^*_c = \mathbf{e}^t_p\Delta t_p \quad (16)$$

Using the approach described in section 2.2.3, the first-order expressions are

$$\Delta\mathbf{r}_b = \Delta\mathbf{Y}^{t,1}_p + \alpha^t_p \times \mathbf{r}^0_b$$

$$\Delta\mathbf{r}_c = \Delta\mathbf{Y}^{t,1}_p + \alpha^t_p \times \mathbf{r}^0_c + \mathbf{e}^t_p\Delta t_p \quad (17)$$

where α^t_p and $\Delta\mathbf{Y}^{t,1}_p$ represent the first order in $\mathbf{\Omega}$ and $\Delta\mathbf{Y}_1$ in eq 14, respectively. After substituting eqs 17 in eqs 12, we obtain

$$M_{1p}\Delta\mathbf{Y}^{t,1}_p + M_{2p}\Delta\mathbf{Y}^{t,1}_p + M_2\mathbf{e}^t_p\Delta t_p = M\Delta\mathbf{Y}^{t,1}_p + M_2\mathbf{e}^t_p\Delta t_p = 0$$

$$M_{1p}\mathbf{G}^0_{1p} \times \Delta\mathbf{Y}^{t1}_p + M_{2p}\mathbf{G}^0_{2p} \times \Delta\mathbf{Y}^{t1}_p + M_{2p}\mathbf{G}^0_{2p} \times \mathbf{e}^t_p\Delta t_p + \mathbf{I}\alpha^t_p = 0 \quad (18)$$

By solving eqs 18 for α^t_p and $\Delta\mathbf{Y}^{t,1}_p$, we get

$$\Delta\mathbf{Y}^{t1}_p = -M_{2p}M^{-1}\mathbf{e}^t_p\Delta t_p$$

$$\alpha^t_p = -\mathbf{I}^{-1}(M_{2p}\mathbf{G}^0_{2p} \times \mathbf{e}^t_p) \quad (19)$$

where all the quantities are defined in eqs 13. Therefore, we can write

$$\frac{\partial\mathbf{r}_b}{\partial t_p} = -\gamma^{t(1)}_p + \alpha^{t(1)}_p \times \mathbf{r}^0_b$$

$$\frac{\partial\mathbf{r}_c}{\partial t_p} = \gamma^{t(2)}_p - \alpha^{t(2)}_p \times \mathbf{r}^0_c \quad (20)$$

where

$$\gamma^{t(i)}_p = (1 - M_{ip}M^{-1})\mathbf{e}^t_p$$

$$\alpha^{t(i)}_p = \mathbf{I}^{-1}(M_{ip}\mathbf{G}^0_{ip} \times \mathbf{e}^t_p) \quad (21)$$

with i = 1 or 2.

Intra and Inter Angles. Using the approach present in section 2.2.3, we obtain the formulas (the details of the mathematical derivations are reported in section S1.1.1 of the Supporting Information):

$$\frac{\partial\mathbf{r}_b}{\partial\psi_p} = \gamma^{\psi(1)}_p + \alpha^{\psi(1)}_p \times \mathbf{r}^0_b$$

$$\frac{\partial\mathbf{r}_c}{\partial\psi_p} = -\gamma^{\psi(2)}_p - \alpha^{\psi(2)}_p \times \mathbf{r}^0_c \quad (22)$$

where

$$\gamma^{\psi(i)}_p = \mathbf{e}^{\psi}_p \times \{(1 - M^{-1}M_{ip})\mathbf{r}^0_p + M^{-1}M_{ip}\mathbf{G}^0_{ip}\}$$

$$\alpha^{\psi(i)}_p = -\mathbf{I}^{-1}\{M_{ip}\mathbf{G}^0_{ip} \times (\mathbf{e}^{\psi}_p \times \mathbf{r}^0_p) + (\mathbf{I}^{-1} - \mathbf{I}_{ip})\mathbf{e}^{\psi}_p\} \quad (23)$$

2.3. Conversion from Internal Coordinates to Cartesian Ones. For small-amplitude conformational dynamics of the proteins, the Taylor expansion of the Cartesian coordinates[44] around a given conformation (usually an energy minimum) corresponding to a set of internal coordinates q is given by

$$\mathbf{r}_a\{q + \Delta q\} = \mathbf{r}_a\{q\} + \sum_i \frac{\partial\mathbf{r}_a}{\partial q_i}\Delta q_i + \frac{1}{2}\sum_{i,j}\frac{\partial^2\mathbf{r}_a}{\partial q_i\partial q_j}\Delta q_i\Delta q_j + O(q^3) \quad (24)$$

This expression should satisfy the Eckart conditions presented in eq 12. The displacement of each internal coordinate Δq_i in the preceding instantaneous conformation is given by

$$\Delta q_i = \frac{\sqrt{2k_BT}}{\omega_k}a_{ik} \qquad (i = 1, 2, ..., N) \quad (25)$$

where a_{ik} is the ith component of the eigenvector for the kth mode of frequency ω_k. From these conditions, the explicit formulas for the coefficients of the linear terms are the same as those derived in section 2.2. Here, we derive an explicit expression for the coefficients of the quadratic terms in eq 24, which we name the **L** matrix, for N bodies. Previously, **L** was derived for a single body[44] in torsional space. The expressions for the diagonal

DOI: 10.1021/acs.jctc.5b00724
J. Chem. Theory Comput. 2015, 11, 5503−5512

elements and off-diagonal elements in TAS are reported in section S1 of the Supporting Information.

2.3.1. Diagonal Elements of the L Matrix. Inter Distance or Bond Lengths. We consider second-order terms in Δt_p. Therefore, we can write

$$\Delta \mathbf{r}_{b,pp} = \Delta \mathbf{Y}_p^{t,2} + \beta_{pp}^t \times \mathbf{r}_b^0 \Delta t_p \Delta t_p + \frac{1}{2}\alpha_p^t \times (\alpha_p^t \times \mathbf{r}_b^0)\Delta t_p \Delta t_p$$

$$\Delta \mathbf{r}_{c,pp} = \Delta \mathbf{Y}_p^{t,2} + \beta_{pp}^t \times \mathbf{r}_c^0 \Delta t_p \Delta t_p + \frac{1}{2}\alpha_p^t \times (\alpha_p^t \times \mathbf{r}_c^0)\Delta t_p \Delta t_p + \alpha_p^t \times \mathbf{e}_p^t \Delta t_p \Delta t_p \quad (26)$$

Using the same approach as that described in section 2.2.3, the expressions for $\Delta \mathbf{Y}_p^{t,2}$ and β_{pp}^t are obtained:

$$\Delta \mathbf{Y}_p^{t,2} = -\frac{M_{2p}\alpha_p^t \times \mathbf{e}_p^t}{M}$$

$$\beta_{pp}^t = -\frac{1}{2}\mathbf{I}^{-1}[\alpha_p^t \mathbf{T} \alpha_p^t + 2M_{2p}\mathbf{G}_{2p}^0 \times (\alpha_t \times \mathbf{e}_p^t)] \quad (27)$$

Here, $\mathbf{T}$ is a third-order tensor whose elements are given by

$$T_{ijk} = -\sum_a \sum_l m_a \mathbf{r}_{ia} \varepsilon_{jkl} \mathbf{r}_{la} \quad (28)$$

where ε_{ijk} is the Levi−Civita symbol,[51] which is defined by

$$\varepsilon_{123} = \varepsilon_{231} = \varepsilon_{312} = 1$$
$$\varepsilon_{132} = \varepsilon_{213} = \varepsilon_{321} = -1$$
$$\text{all others, } \varepsilon_{ijk} = 0 \quad (29)$$

Note that the j components of $\alpha_p^t \mathbf{T} \alpha_p^t$ are given by $\sum_{ik} \alpha_{pi}^t T_{ijk} \alpha_{pk}^t$. The final form for the diagonal elements is

$$\frac{\partial^2 \mathbf{r}_b}{\partial t_p^2} = \zeta_{pp}^{t(1)} + \beta_{pp}^{t(1)} \times \mathbf{r}_b^0 + \frac{1}{2}\alpha_p^{t(1)} \times (\alpha_p^{t(1)} \times \mathbf{r}_b^0)$$

$$\frac{\partial^2 \mathbf{r}_c}{\partial t_p^2} = \zeta_{pp}^{t(2)} + \beta_{pp}^{t(2)} \times \mathbf{r}_c^0 + \frac{1}{2}\alpha_p^{t(2)} \times (\alpha_p^{t(2)} \times \mathbf{r}_c^0) \quad (30)$$

where

$$\zeta_{pp}^{t(i)} = (M^{-1}M_{ip} - 1)\alpha_p^{t(i)} \times \mathbf{e}_p^t$$

$$\beta_{pp}^{t(i)} = -\frac{1}{2}\mathbf{I}^{-1}[\alpha_p^{(i)t} \mathbf{T} \alpha_p^{(i)t} - 2M_{ip}\mathbf{G}_{ip}^0 \times (\alpha_p^{t(i)} \times \mathbf{e}_p^t)] \quad (31)$$

2.3.2. Off-Diagonal Elements of the L Matrix. Inter/Intra Distances. We consider the case of two translations. In this case, the molecule can be regarded as three bodies connected by two bonds p and q (see Figure 3). First, we translate rigid body 3 along the unit vector $\mathbf{e}_q^t$ by an infinitesimal quantity Δt_q:

$$\Delta \mathbf{r}_d^{**} = \mathbf{e}_q^t \Delta t_q \quad (32)$$

Then, we translate rigid bodies 2 and 3 along the unit vector $\mathbf{e}_p^t$ by an infinitesimal quantity Δt_p:

$$\Delta \mathbf{r}_c^{*} = \mathbf{e}_t^t \Delta t_p$$

$$\Delta \mathbf{r}_d^{*} = \Delta \mathbf{r}_d^{**} + \mathbf{e}_p^t \Delta t_p \quad (33)$$

Using the approach present in section 2.2.3 and by considering only the second-order terms $\Delta t_q \Delta t_p$, we can obtain the defining

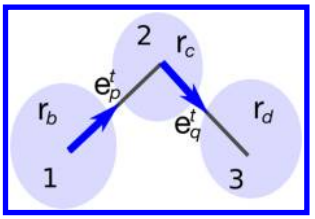


Figure 3. Schematic of a molecule regarded as three rigid bodies 1, 2, and 3 connected by two chemical bonds p and q for the translation t_{12} and t_{23}. $\mathbf{e}_p$ and $\mathbf{e}_q$ are the unit vectors in the direction of bond p and q, respectively. The position vectors of the atoms belonging to rigid bodies 1, 2, and 3 are denoted by $\mathbf{r}_b$, $\mathbf{r}_c$, and $\mathbf{r}_d$, respectively.

formulas. (The details of the mathematical derivations are reported in section S1.3.1 of the Supporting Information.) Finally, we obtain

$$\frac{\partial^2 \mathbf{r}_b}{\partial t_p \partial t_q} = \zeta_{pq}^{t(1)} + \beta_{pq}^{t(1)} \times \mathbf{r}_b^0 + \frac{1}{2}\alpha_p^{t(1)} \times (\alpha_q^{t(1)} \times \mathbf{r}_b^0) + \frac{1}{2}\alpha_q^{t(1)} \times (\alpha_p^{t(1)} \times \mathbf{r}_b^0)$$

$$\frac{\partial^2 \mathbf{r}_c}{\partial t_p \partial t_q} = \zeta_{pq}^{t(2)} + \beta_{pq}^{t(2)} \times \mathbf{r}_c^0 - \frac{1}{2}[\alpha_p^{t(2)} \times (\alpha_q^{t(1)} \times \mathbf{r}_c^0) + \alpha_q^{t(1)} \times (\alpha_p^{t(2)} \times \mathbf{r}_c^0)]$$

$$\frac{\partial^2 \mathbf{r}_d}{\partial t_p \partial t_q} = \zeta_{pq}^{t(3)} + \beta_{pq}^{t(3)} \times \mathbf{r}_d^0 + \frac{1}{2}\alpha_p^{t(2)} \times (\alpha_q^{t(2)} \times \mathbf{r}_d^0) + \frac{1}{2}\alpha_q^{t(2)} \times (\alpha_p^{t(2)} \times \mathbf{r}_d^0) \quad (34)$$

Here,

$$\zeta_{pq}^{t(1)} = -[(1 - M_{1p}M^{-1})\alpha_q^{t,(1)} \times \mathbf{e}_p^t + (1 - M_{1q}M^{-1})\alpha_p^{t,(1)} \times \mathbf{e}_q^t]$$
$$\zeta_{pq}^{t(2)} = [(1 - M_{2p}M^{-1})\alpha_q^{t,(1)} \times \mathbf{e}_p^t + (1 - M_{1q}M^{-1})\alpha_p^{t,(2)} \times \mathbf{e}_q^t]$$
$$\zeta_{pq}^{t(3)} = -[(1 - M_{2p}M^{-1})\alpha_q^{t,(2)} \times \mathbf{e}_p^t + (1 - M_{2q}M^{-1})\alpha_p^{t,(2)} \times \mathbf{e}_q^t] \quad (35)$$

$$\beta_{pq}^{t(1)} = -\frac{1}{2}\mathbf{I}^{-1}[\alpha_p^{t,(1)} \mathbf{T} \alpha_q^{t,(1)} + \alpha_q^{t,(1)} \mathbf{T} \alpha_p^{t,(1)} - 2M_{1p}\mathbf{G}_{1p} \times (\alpha_q^{t,(1)} \times \mathbf{e}_p^t) - 2M_{1q}\mathbf{G}_{1q} \times (\alpha_p^{t,(1)} \times \mathbf{e}_q^t)]$$

$$\beta_{pq}^{t(2)} = \frac{1}{2}\mathbf{I}^{-1}[\alpha_p^{t,(2)} \mathbf{T} \alpha_q^{t,(1)} + \alpha_q^{t,(1)} \mathbf{T} \alpha_p^{t,(2)} - 2M_{2p}\mathbf{G}_{2p} \times (\alpha_q^{t,(1)} \times \mathbf{e}_p^t) - 2M_{1q}\mathbf{G}_{1q} \times (\alpha_p^{t,(2)} \times \mathbf{e}_q^t)]$$

$$\beta_{pq}^{t(3)} = -\frac{1}{2}\mathbf{I}^{-1}[\alpha_p^{t,(2)} \mathbf{T} \alpha_q^{t,(2)} + \alpha_q^{t,(2)} \mathbf{T} \alpha_p^{t,(2)} - 2M_{2p}\mathbf{G}_{2p} \times (\alpha_q^{t,(2)} \times \mathbf{e}_p^t) - 2M_{2q}\mathbf{G}_{2q} \times (\alpha_p^{t,(2)} \times \mathbf{e}_q^t)] \quad (36)$$

Inter/Intra Angles and Inter/Intra Distances. We next consider the case of a translation and a rotation. In this case, the molecule can be regarded as three bodies connected by two bonds p and q. Using the approach presented in section 2.2.3 and by considering only the second-order terms $\Delta t_q \Delta \psi_p$, we can obtain (the details of the mathematical derivations are reported

DOI: 10.1021/acs.jctc.5b00724
J. Chem. Theory Comput. 2015, 11, 5503−5512

in section S1.3.2 of the Supporting Information):

$$\frac{\partial^2 \mathbf{r}_b}{\partial \psi_p \partial t_q} = \zeta_{pq}^{\psi,t(1)} + \beta_{pq}^{\psi,t(1)} \times \mathbf{r}_b^0 + \frac{1}{2}\alpha_p^{\psi(1)} \times (\alpha_q^{t(1)} \times \mathbf{r}_b^0) + \frac{1}{2}\alpha_q^{t(1)} \times (\alpha_p^{\psi(1)} \times \mathbf{r}_b^0)$$

$$\frac{\partial^2 \mathbf{r}_c}{\partial \psi_p \partial t_q} = \zeta_{pq}^{\psi,t(2)} + \beta_{pq}^{\psi,t(2)} \times \mathbf{r}_c^0 - \frac{1}{2}\{\alpha_q^{\psi(2)} \times (\alpha_q^{t(1)} \times \mathbf{r}_c^0) + \alpha_q^{t(1)} \times (\alpha_p^{\psi(2)} \times \mathbf{r}_c^0)\}$$

$$\frac{\partial^2 \mathbf{r}_d}{\partial \psi_p \partial t_q} = \zeta_{pq}^{\psi,t(3)} + \beta_{pq}^{\psi,t(3)} \times \mathbf{r}_d^0 + \frac{1}{2}\alpha_p^{\psi(2)} \times (\alpha_q^{t(2)} \times \mathbf{r}_d^0) + \frac{1}{2}\alpha_q^{t(2)} \times (\alpha_p^{\psi(2)} \times \mathbf{r}_d^0) \quad (37)$$

Here,

$$\zeta_{pq}^{\psi,t(1)} = M^{-1}\{M_{1p}[\alpha_q^{(1)t} \times (\mathbf{e}_p^{\psi} \times \mathbf{G}_{1p})] + (M - M_{1p}) \times [\alpha_q^{(1)t} \times (\mathbf{e}_p^{\psi} \times \mathbf{r}_p^0)] - (M - M_{1q})[(\alpha_q^{(1)\psi} + \mathbf{e}_p^{\psi}) \times \mathbf{e}_q^t]\}$$

$$\zeta_{pq}^{\psi,t(2)} = -M^{-1}\{M_{2p}[\alpha_q^{(1)t} \times (\mathbf{e}_p^{\psi} \times \mathbf{G}_{2p})] + (M - M_{2p}) \times [\alpha_q^{(1)t} \times (\mathbf{e}_p^{\psi} \times \mathbf{r}_p^0)] - (M - M_{1q})[(\alpha_q^{(2)\psi} \times \mathbf{e}_q^t)]\}$$

$$\zeta_{pq}^{\psi,t(3)} = M^{-1}\{M_{2p}[\alpha_q^{(2)t} \times (\mathbf{e}_p^{\psi} \times \mathbf{G}_{2p})] + (M - M_{2p}) \times [\alpha_q^{(2)t} \times (\mathbf{e}_p^{\psi} \times \mathbf{r}_p^0)] - (M - M_{2q})(\alpha_p^{(2)\psi} \times \mathbf{e}_q^t)\} \quad (38)$$

$$\beta_{pq}^{\psi,t(1)} = -\frac{1}{2}\mathbf{I}^{-1}\{\alpha_p^{\psi(1)}\mathbf{T}\alpha_q^{t(1)} + \alpha_q^{t(1)}\mathbf{T}\alpha_p^{\psi(1)} + 2\alpha_q^{t(1)}(\mathbf{T} - \mathbf{T}_{1p})\mathbf{e}_p^{\psi} + 2M_{1p}\mathbf{G}_{1p} \times [\alpha_q^{t(1)} \times (\mathbf{e}_p^{\psi} \times \mathbf{r}_p^0)] - 2M_{1q}\mathbf{G}_{1q} \times [(\alpha_p^{\psi(1)} + \mathbf{e}_p^{\psi}) \times \mathbf{e}_q^t]\}$$

$$\beta_{pq}^{\psi,t(2)} = \frac{1}{2}\mathbf{I}^{-1}\{\alpha_p^{\psi(2)}\mathbf{T}\alpha_q^{t(1)} + \alpha_q^{t(1)}\mathbf{T}\alpha_p^{\psi(2)} + 2\alpha_q^{t(1)}(\mathbf{T} - \mathbf{T}_{2p})\mathbf{e}_p^{\psi} + 2M_{2p}\mathbf{G}_{2p} \times [\alpha_q^{t(1)} \times (\mathbf{e}_p^{\psi} \times \mathbf{r}_p^0)] - 2M_{1q}\mathbf{G}_{1q} \times [(\alpha_p^{\psi(2)} \times \mathbf{e}_q^t)]\}$$

$$\beta_{pq}^{\psi,t(3)} = -\frac{1}{2}\mathbf{I}^{-1}\{\alpha_p^{\psi(2)}\mathbf{T}\alpha_q^{t(2)} + \alpha_q^{t(2)}\mathbf{T}\alpha_p^{\psi(2)} + 2\alpha_q^{t(2)}(\mathbf{T} - \mathbf{T}_{2p})\mathbf{e}_p^{\psi} + 2M_{2p}\mathbf{G}_{2p} \times [\alpha_q^{t(2)} \times (\mathbf{e}_p^{\psi} \times \mathbf{r}_p^0)] - 2M_{2q}\mathbf{G}_{2q} \times [(\alpha_p^{\psi(2)} \times \mathbf{e}_q^t)]\} \quad (39)$$

We can also start by rotating rigid body 3 along the unit vector $\mathbf{e}_q^{\psi}$ and then translate rigid bodies 2 and 3 along the unit vector $\mathbf{e}_p^t$. The details of the mathematical derivations and the formulas are reported in section S1.3.2 of the Supporting Information.

2.3.3. Average Properties. Because of the conversion from ICS to CCS, we can calculate two average quantities: the displacements of mean atomic positions from their positions in the minimum-energy conformation, and the mean square fluctuation of each atom from its displaced mean position. Because our NMA is performed in ICS, the time behavior of the internal variables (q_i) is given by a linear combination of the normal modes in ICS and the resulting conformation should be the minimum-energy conformation. However, if we consider the terms up to second order or more (see eq 24), the conversion from ICS to CCS is no longer linear. The resulting nonlinearity produces a displacement of atomic positions. The average over all conformations is given by

$$\langle \Delta \mathbf{r}_a \rangle = \frac{1}{2}\sum_{i,j} L_{aij}\langle \delta q_i \delta q_j \rangle = \frac{1}{4}\sum_{i,j} L_{aij} \sum_k A_{ik}A_{jk}\alpha_k^2 \quad (40)$$

and the mean fluctuations are given by

$$\langle (\Delta \mathbf{r}_a)^2 \rangle = \left\langle \left(\sum_i K_{ai}\delta q_i + \frac{1}{2}\sum_{i,j} L_{aij}\delta q_i \delta q_j \right)^2 \right\rangle = \frac{1}{2}\sum_{ij} K_{ai}K_{aj} \sum_k A_{ik}A_{jk}\alpha_k^2 + \frac{3}{32}\sum_{ijlt} L_{aij}L_{alt} \sum_k A_{ik}A_{jk}A_{lk}A_{tk}\alpha_k^4 \quad (41)$$

3. COMPUTATIONAL DETAILS

3.1. Hessian, Kinetic, and L Matrices. As described in section 2.2, to perform an energy minimization, we choose a reference body as fixed in space; this choice implies that its rotations (τ and θ) are projected to the bodies involved in the relative rotation. When coupling intra and inter variables together, one must avoid moving the pivot atoms of the bodies or changing their relative orientations: To overcome this problem, we must define the rigid bodies for each case and the direction of the unit vector describing the intra rotations; in this way, the pivot atoms remain fixed. A similar approach must be taken when the kinetic and **L** matrices are computed. Figure 4 shows an example of definition of rigid bodies, represented as linear chains, when we compute the kinetic matrix for a rotation preceding or following the pivot atom in a system of two bodies. In section S2.1 of the Supporting Information, we show the approach for other combinations of intra variables, or combinations of intra and inter variables within an ensemble of N proteins.

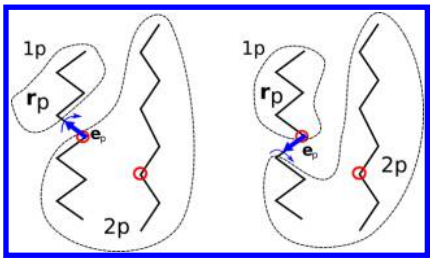


Figure 4. Schematic representation of a torsion in the first protein preceding (left) or following (right) the pivot atom (red circle). The system is treated as two bodies, 1p and 2p, connected by a chemical bond p. Relative rotation involves a torsion angle around p. $\mathbf{r}_p$ are the position vectors of the center of mass of rigid body 1 and of the atom at an end of bond p in rigid body 1. $\mathbf{e}_p$ is the unit vector in the direction of bond p. The red circles indicate the pivot atom of each protein.

3.2. PaLaCe Coarse-Grain Model. PaLaCe[48] (named after the developers Pasi, Lavery, and Ceres) is a CG protein model developed in our group that can be used to simulate the behavior of individual proteins or of protein complexes. PaLaCe uses a two-tier representation of polypeptide chains. The first tier is used for nonbonded interactions, where each residue is represented by 1−3 pseudo-atoms (that is, beads representing groups of atoms) using the Zacharias representation.[52] In this model, each amino acid is represented by one pseudo-atom located at the Cα position, and either one or two pseudo-atoms (SC1, SC2) representing the side chain (with the exception of Gly). Ala, Ser, Thr, Val, Leu, Ile, Asn, Asp, and Cys have a single pseudo-atom located at the geometrical center of the side-chain heavy atoms. For the remaining amino acids, a first pseudo-atom

DOI: 10.1021/acs.jctc.5b00724
J. Chem. Theory Comput. 2015, 11, 5503−5512

is located midway between the Cβ and Cγ atoms, whereas the second is placed at the geometrical center of the remaining side-chain heavy atoms. In this way, residue specificity is maintained at a relatively low computational cost. The second tier is used for bonded interactions and for backbone hydrogen bonds. It involves atomic beads for N, C′, and Cα. For hydrogen bonding, the backbone oxygen and amide hydrogen positions are geometrically constructed on the basis of Cα, N, and C′ positions, avoiding any additional degrees of freedom. Solvent interactions are addressed using an implicit model based on residue burial. This further accelerates the calculations. The functions used to calculate the potential energy of the protein(s) are given in section S2.2 of the Supporting Information.

3.3. Calculation Protocol. PALINCOM (PALace Internal COordinate Minimization) was developed to energy minimize protein conformations in ICS using PaLaCe Force Field, but it can also perform energy minimization in CCS; both approaches use a modified Newton minimizer (Harwell VA13A). After energy minimization, we performed NMA in ICS using PaLaCe (hereafter termed iNMA-P) and in CCS using PaLaCe (hereafter termed cNMA-P). For further comparisons, we also used an elastic network model limited to Cα nodes to obtain ANM normal modes (reference state = pdb structure, cutoff = 9 Å, force constant = 0.6 kcal mol^{-1}) (hereafter termed cNMA-A). In each case studied, all the vibrational frequencies obtained were positive (with the exclusion of the six zero-frequency modes representing overall translation and rotation with cNMA calculations). In the case of iNMA, for the first 20 modes, i.e., for a set of $\{\Delta q_i\}$ given by eq 25, using eq 24, we calculated $\Delta \mathbf{r}_a$ to perform the conversion from ICS to CCS.

We tested the different NMA calculations by studying their ability to represent unbound-to-bound conformational changes of proteins that form binary complexes. In order to quantify the results, we calculate O_j, the overlap between the conformational change predicted by the jth normal mode and the observed unbound-to-bound conformational change:[53,54]

$$O_j = \frac{\sum_{i=1}^{N} \mathbf{a}_{ij} \cdot \Delta \mathbf{r}_i}{(\sum_{i=1}^{N} \mathbf{a}_{ij}^2 \sum_{i=1}^{N} \Delta \mathbf{r}_i^2)^{1/2}} \tag{42}$$

where $\Delta \mathbf{r}_i$ is the vector describing the unbound-to-bound conformational change for an atom i and $\mathbf{a}_{ij}$ represents the ith displacement in the jth mode. For iNMA, $\mathbf{a}_{ij}$ is obtained from the ICS to CCS conversion. To compute the vector $\Delta \mathbf{r}_i$, we must align the unbound and bound structures. This was performed using the mass-weighted version of the McLachlan algorithm,[55] as implemented in the program ProFit for cNMA and iNMA, respectively. We also computed the cumulative overlap, O, which is defined by

$$O = (\sum_j O_j^2)^{1/2} \tag{43}$$

4. RESULTS

We performed calculations for 61 complexes (122 single proteins) taken from Protein Docking Benchmark version 4.0[56] and present in the Protein Data Bank (PDB)[57] (the list of PDB codes is given in section S3 of the Supporting Information); we considered the same subset of residues for the bound and unbound structures.

As a preliminary test, we compared the root-mean-square fluctuations (RMSFs) per residue of the chemotaxis protein CheY, PDB code 1TMY, using eq 41, with those derived from

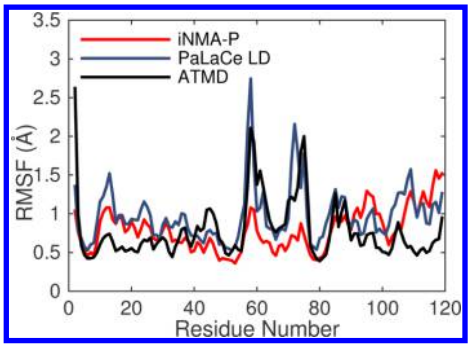


Figure 5. Comparison of iNMA-P backbone fluctuations (red) with those obtained from PaLaCe LD (blue) and ATMD (black) computed for chemotaxis protein CheY, PDB code 1TMY. In each case, the RMSF value (in Å) is computed considering Cα, N, and C′ atoms of the backbone.

trajectories using atomistic molecular dynamic simulations in explicit solvent (ATMD) and PaLaCe Langevin dynamic simulations in implicit solvent (PaLaCe MD) (see Figure 5). The lengths of the trajectories were fixed at 100 ns. Although we would not expect the iNMA-P-derived fluctuations to exactly reproduce the all-atom MD results, the peak positions correspond reasonably well. Multiplying the NMA data by a factor of 1.4[58] also leads to a reasonably good agreement in magnitude, with the exception of region around residues 58 and 76 (corresponding to loops), despite that fact that our analysis is limited to torsion angles and includes only simplified electrostatic and solvent terms.

We now consider the extent to which different normal-mode calculations are able to model the changes between unbound and bound protein conformations. Figure 6 shows the results for

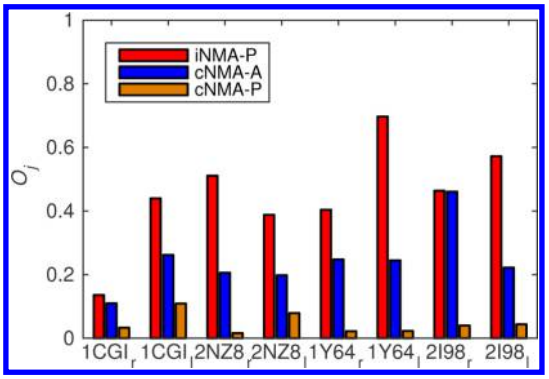


Figure 6. Maximum overlap, O_j, for four complexes (PDB code: 1CGI, 2NZ8, 1Y64, 2I98) involving large conformational changes upon binding (iRMSD > 2 Å). Overlaps are calculated using the 20 lowest modes obtained with iNMA-P (red), cNMA-A (blue), and cNMA-P (orange). The indexes l and r represent the first and second PDB id reported in section S3 of the Supporting Information.

proteins belonging to four binary complexes that undergo large conformational rearrangements upon binding, as indicated by RMSD values for the heavy atoms forming a protein−protein interface (iRMSD) of >2 Å. As a metric, we use the maximum overlap O_j (unbound-to-bound) computed for the 20 lowest frequency modes obtained by iNMA-P, cNMA-P, and cNMA-A. First, we observe that, in the case of PaLaCe Cartesian coordinate modes, the maximum overlap is very low, compared to the other two methods. Although, in a single case, the maximum overlap is similar for iNMA-P and cNMA-P, overall, the predictions of

DOI: 10.1021/acs.jctc.5b00724
J. Chem. Theory Comput. 2015, 11, 5503−5512

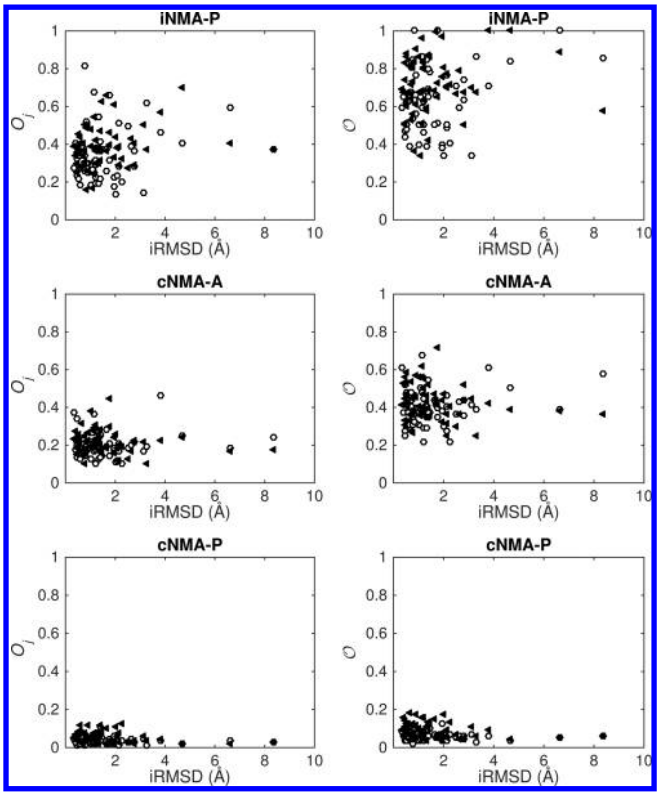


Figure 7. Maximum overlap, O_j (left) and cumulative overlap O (right) for 122 proteins, as a function of iRMSD, computed using the 20 lowest modes obtained with iNMA-P (top), cNMA-A (middle), and cNMA-P (bottom). Legend: (○) PDB id 1, (◀) PDB id 2. The PDB code of the complex, iRMSD, PDB id 1, and PDB id 2 are reported in Table S1 of the Supporting Information.

conformational change using iNMA-P are substantially better either of the other two approaches.

Figure 7 summarizes the maximum, O_j, and the cumulative overlap, O, as a function of iRMSD (see Table S1 in the Supporting Information) for the lowest 20 modes of a full set of 122 proteins that we have studied using the iNMA-P, cNMA-A, and cNMA-P approaches. Considering first the maximum overlap, we observe a range of values of 0.18 → 0.82 for iNMA-P calculations versus only 0.10 → 0.46 for cNMA-A and 0.02 → 0.12 for cNMA-P. Note also that the iNMA-P approach continues to provide good results, even for proteins that undergo large conformational changes (iRMSD > 2 Å). We remark that the overall trend of the results for cNMA-A is not affected by the parameters chosen for the elastic network, but the maximum overlap decreases as the cutoff or force constant increases.

Because the normal modes represent an orthonormal space, we can easily combined them and compute the cumulative overlap. Once again, we see that iNMA-P is able to better describe the transition from unbound to bound structures and, for some proteins, we reach a cumulative overlap equal to 1 (i.e., a complete description of the transition unbound-to-bound conformation) including for proteins undergoing large conformational changes. Figure 8 shows an example of the conformational change of an unbound structure (Actin protein, PDB code 1IJI) upon the application of the lowest mode, as calculated with iNMA-P. The resulting conformation is closely approaches the bound confirmation; only 4% of the residues show significant torsional changes, but these residues account for 50% of the overall motion. We would like to stress that the knowledge of the key residues

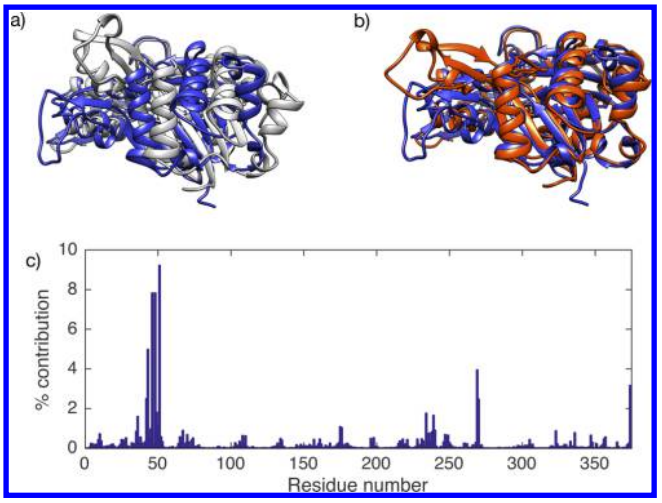


Figure 8. Example of the movement due to the lowest mode, as calculated by iNMA-P for Actin protein, PDB code 1IJI. (a) Superimposition of the unbound (gray) and bound structures (blue, PDB code: 2BTF, chain A). (b) Superimposition of the bound structure and the structure obtained by applying the lowest mode to the unbound form (orange). (c) Percentage contribution of the torsional variables in the lowest mode, as a function of the residue number.

responsible for the conformational change is an intrinsic feature of iNMA, but is difficult to determine using cNMA. Furthermore, thanks to the transformation from ICS to CCS, we can also identify the location of the largest structural displacements due to the key torsional variables for each mode.

The iNMA-P results confirm that unbound protein structures can often be perturbed along a single low-frequency mode to produce a conformation that is close to the bound one. This behavior can advantageously be used in docking algorithms to account for protein flexibility in an efficient way; however, the problem of choosing the "correct" mode(s) remains. To attempt to solve this problem, Figures 9 and 10 illustrate the unbound-to-

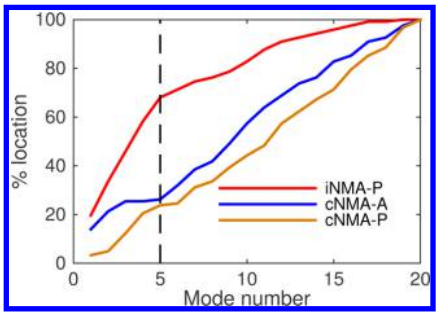


Figure 9. Percentage location of the maximum overlap, O_j, averaged over the 122 proteins studied, for the 20 lowest frequency modes: iNMA-P (red), cNMA-A (blue), cNMA-P (orange). The dashed black line represents the percentage overlap accumulated for only the first five modes.

bound overlap among the lowest-frequency modes for the 122 proteins studied. Figure 9 presents the cumulative results for the first 20 modes, while Figure 10 shows the results mode by mode. Both figures show that iNMA-P performs better in concentrating the optimal modes for predicting conformational change at the lowest frequencies. Indeed, while the cNMA-A approach only places the maximum overlap within the five lowest modes in ~30% of the cases, independent of the choice of the spring force constant or distance cutoff,[29,33] the iNMA-P approach succeeds

DOI: 10.1021/acs.jctc.5b00724
J. Chem. Theory Comput. 2015, 11, 5503−5512

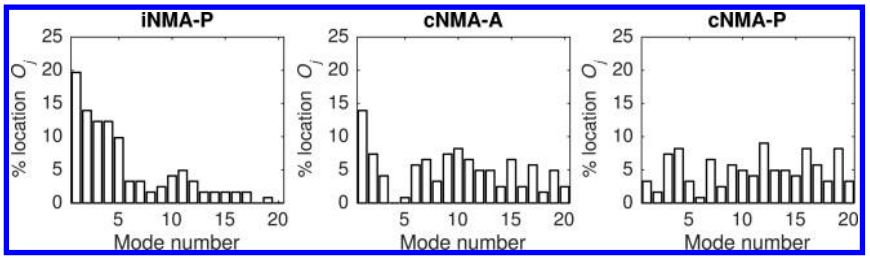


Figure 10. Percentage location of the maximum overlap, O_j, averaged over the 122 proteins studied for the lowest 20 modes: iNMA-P (left), cNMA-A (middle), cNMA-P (right).

in 70% of cases and, moreover, shows a monotonically increasing probability of finding the maximum overlap as the frequency decreases within the first five modes. The cNMA-A and cNMA-P distributions are considerably broader and we can note significant percentage probabilities for maximum overlap modes, even beyond the first 10. We conclude that iNMA is likely to be able to model even large conformational changes with only 5−10 low-frequency modes, simplifying the choice of representing conformational flexibility in NMA space during protein docking.

5. CONCLUSIONS

In this work, we have described the implementation of normal mode analysis in internal coordinate space (iNMA) for both individual proteins and for protein complexes. We have generalized the approach proposed by Sunada and Gō[44] to determine the Cartesian coordinate movements corresponding to the internal coordinate the normal modes for N bodies using a second-order approximation.

We have applied iNMA to predictions of the unbound-to-bound conformational changes occurring in proteins when they form binary complexes. These tests were performed in 122 proteins that constitute 61 binary complexes using two quantitative metrics: the overlap O_j per mode j and the cumulative overlap, $\bar{O}$, for the 20 lowest-frequency modes. Using the PaLaCe coarse-grain (CG) model and comparing internal (iNMA) with Cartesian coordinate (cNMA) modes shows that the internal coordinate approach is clearly superior. The iNMA approach generally requires fewer modes to represent the conformational transitions and it makes good predictions even for large-scale transitions. The most relevant modes are found within the five lowest frequency modes in 70% of the cases studied and, within this subset, the probability of finding the single optimal mode increases monotonically with decreasing frequency. The cumulative overlap of the first 20 modes can reach unity (i.e., a perfect description of the unbound-to-bound transition). We have also made a comparison of iNMA modes using a simple elastic network representation (ANM) rather than the PaLaCe force field and find that results are not surprisingly better, with the more-detailed representation and the more-sophisticated force field provided by PaLaCe.

These promising results open a new route to treating conformational flexibility during the formation of protein−protein interactions. By adding a small number of iNMA low-frequency modes to the rigid-body degrees of freedom typically used in protein docking methodologies, it should be possible to represent changes in internal conformations with little additional computational expense.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jctc.5b00724.

The derivation of the diagonal and off-diagonal elements of **L** matrix for torsions is presented. The definition of rigid bodies for the Hessian, kinetic, and L matrices is explained. The PaLaCe force field is described, and the data related to the proteins under investigation are reported (PDF)

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: richard.lavery@ibcp.fr.

Author Contributions

The manuscript was written with contributions from both authors. Both authors have worked on and approved the final version of the manuscript.

Funding

The European Union Seventh Framework Programme (No. FP7/2007−2013), under Grant Agreement No. 604102 − The Human Brain Project (Subproject 6, Brain Simulation, WP6.3: Molecular Dynamics Simulation, T6.3.1: Atomistic and coarse-grain model simulations), and the ANR "Future Investments in Bioinformatics" program (through the project MAPPING (ANR-11-BINF-0003)) are acknowledged for funding this research.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors thank the French supercomputer CINES for a generous allocation of computer time and Laura Pontille for help in testing the NMA algorithms.

■ ABBREVIATIONS

ANM, Anisotropic Network Model; CCS, Cartesian Coordinate Space; CG, coarse-grain; cNMA, Cartesian Normal Mode Analysis; ICS, internal coordinate space; iNMA, internal coordinate normal mode analysis; iRMSD, interface root-mean-square deviation; RMSD, root-mean-square deviation; TAS, torsional angular space

■ REFERENCES

(1) Alberts, B. *Cell* **1998**, *92*, 291−294.
(2) Minton, A. P. *Curr. Opin. Struct. Biol.* **2000**, *10*, 34−39.
(3) Ellis, R. J.; Minton, A. P. *Nature* **2003**, *425*, 27−28.
(4) Deeds, E. J.; Ashenberg, O.; Gerardin, J.; Shakhnovich, E. I. *Proc. Natl. Acad. Sci. U. S. A.* **2007**, *104*, 14952−14957.
(5) Stein, A.; Mosca, R.; Aloy, P. *Curr. Opin. Struct. Biol.* **2011**, *21*, 200−208.

DOI: 10.1021/acs.jctc.5b00724
J. Chem. Theory Comput. 2015, 11, 5503−5512

(6) Garma, L.; Mukherjee, S.; Mitra, P.; Zhang, Y. *PLoS One* **2012**, *7*, e38913.
(7) Aloy, P.; Pichaud, M.; Russell, R. B. *Curr. Opin. Struct. Biol.* **2005**, *15*, 15−22.
(8) Vajda, S.; Kozakov, D. *Curr. Opin. Struct. Biol.* **2009**, *19*, 164−170.
(9) Huang, S. Y. *Drug Discovery Today* **2014**, *19*, 1081−1096.
(10) Betts, M. J.; Sternberg, M. J. *Protein Eng., Des. Sel.* **1999**, *12*, 271−283.
(11) Huber, R.; Bennett, W. S. *Biopolymers* **1983**, *22*, 261−279.
(12) Levitt, M.; Sander, C.; Stern, P. S. *J. Mol. Biol.* **1985**, *181*, 423−447.
(13) Kitao, A.; Gō, N. *J. Comput. Chem.* **1991**, *12*, 359−368.
(14) Tirion, M. M. *Phys. Rev. Lett.* **1996**, *77*, 1905.
(15) Bahar, I.; Lezon, T. R.; Bakan, A.; Shrivastava, I. H. *Chem. Rev.* **2010**, *110*, 1463−1497.
(16) Orozco, M. *Chem. Soc. Rev.* **2014**, *43*, 5051−5066.
(17) Doruker, P.; Atilgan, A. R.; Bahar, I. *Proteins: Struct., Funct., Genet.* **2000**, *40*, 512−524.
(18) Atilgan, A. R.; Durell, S. R.; Jernigan, R. L.; Demirel, M. C.; Keskin, O.; Bahar, I. *Biophys. J.* **2001**, *80*, 505−515.
(19) Tama, F.; Sanejouand, Y.-H. *Protein Eng., Des. Sel.* **2001**, *14*, 1−6.
(20) Micheletti, C.; Carloni, P.; Maritan, A. *Proteins: Struct., Funct., Genet.* **2004**, *55*, 635−645.
(21) Kurkcuoglu, O.; Jernigan, R. L.; Doruker, P. *Biochemistry* **2006**, *45*, 1173−1182.
(22) Frappier, V.; Najmanovich, R. J. *PLoS Comput. Biol.* **2014**, *10*, e1003569.
(23) Zheng, W.; Brooks, B. R.; Thirumalai, D. *Proc. Natl. Acad. Sci. U. S. A.* **2006**, *103*, 7664−7669.
(24) Bahar, I.; Rader, A. J. *Curr. Opin. Struct. Biol.* **2005**, *15*, 586−592.
(25) Maragakis, P.; Karplus, M. *J. Mol. Biol.* **2005**, *352*, 807−822.
(26) Jensen, F.; Palmer, D. S. *J. Chem. Theory Comput.* **2011**, *7*, 223−230.
(27) Palmer, D. S.; Jensen, F. *Proteins: Struct., Funct., Genet.* **2011**, *79*, 2778−2793.
(28) Tobi, D.; Bahar, I. *Proc. Natl. Acad. Sci. U. S. A.* **2005**, *102*, 18908−18913.
(29) Dobbins, S. E.; Lesk, V. I.; Sternberg, M. J. *Proc. Natl. Acad. Sci. U. S. A.* **2008**, *105*, 10390−10395.
(30) May, A.; Zacharias, M. *Proteins: Struct., Funct., Genet.* **2008**, *70*, 794−809.
(31) Andrusier, N.; Mashiach, E.; Nussinov, R.; Wolfson, H. J. *Proteins: Struct., Funct., Genet.* **2008**, *73*, 271−289.
(32) Zacharias, M. *Curr. Opin. Struct. Biol.* **2010**, *20*, 180−186.
(33) Stein, A.; Rueda, M.; Panjkovich, A.; Orozco, M.; Aloy, P. *Structure* **2011**, *19*, 881−889.
(34) Bruccoleri, R. E.; Karplus, M.; McCammon, J. A. *Biopolymers* **1986**, *25*, 1767−1802.
(35) Brooks, B.; Karplus, M. *Proc. Natl. Acad. Sci. U. S. A.* **1985**, *82*, 4995−4999.
(36) Trakhanov, S.; Vyas, N. K.; Luecke, H.; Kristensen, D. M.; Ma, J.; Quiocho, F. A. *Biochemistry* **2005**, *44*, 6597−6608.
(37) Mahajan, S.; Sanejouand, Y. H. *Arch. Biochem. Biophys.* **2015**, *567*, 59−65.
(38) Ma, J. *Curr. Protein Pept. Sci.* **2004**, *5*, 119.
(39) Ma, J. *Structure* **2005**, *13*, 373−380.
(40) Kitao, A.; Hayward, S.; Gō, N. *Biophys. Chem.* **1994**, *52*, 107−114.
(41) Braun, W.; Yoshioki, S.; Gō, N. *J. Phys. Soc. Jpn.* **1984**, *53*, 3269−3275.
(42) Higo, J.; Seno, Y.; Gō, N. *J. Phys. Soc. Jpn.* **1985**, *54*, 4053−4058.
(43) Ishida, H.; Jochi, Y.; Kidera, A. *Proteins: Struct., Funct., Genet.* **1998**, *32*, 324−333.
(44) Sunada, S.; Gō, N. *J. Comput. Chem.* **1995**, *16*, 328−336.
(45) Mendez, R.; Bastolla, U. *Phys. Rev. Lett.* **2010**, *104*, 228103.
(46) Bray, J. K.; Weiss, D. R.; Levitt, M. *Biophys. J.* **2011**, *101*, 2966−2969.
(47) Noguti, T.; Gō, N. *J. Phys. Soc. Jpn.* **1983**, *52*, 3283−3288.
(48) Pasi, M.; Lavery, R.; Ceres, N. *J. Chem. Theory Comput.* **2013**, *9*, 785−793.
(49) Goldstein, H.; Poole, C. P.; Safko, J. L. *Classical Mechanics*; Addison Wesley: San Francisco, CA, 2002.
(50) Eckart, C. *Phys. Rev.* **1935**, *47*, 552.
(51) Arfken, G. B. *Mathematical Methods for Physicists*; Academic Press: New York, 2013.
(52) Zacharias, M. *Protein Sci.* **2003**, *12*, 1271−1282.
(53) Marques, O.; Sanejouand, Y. H. *Proteins: Struct., Funct., Genet.* **1995**, *23*, 557−560.
(54) Ma, J.; Karplus, M. *J. Mol. Biol.* **1997**, *274*, 114−131.
(55) McLachlan, A. D. *Acta Crystallogr., Sect. A: Cryst. Phys., Diffr., Theor. Gen. Crystallogr.* **1982**, *38*, 871−873.
(56) Hwang, H.; Vreven, T.; Janin, J.; Weng, Z. *Proteins: Struct., Funct., Genet.* **2010**, *78*, 3111−3114.
(57) Berman, H. M.; Battistuz, T.; Bhat, T. N.; Bluhm, W. F.; Bourne, P. E.; Burkhardt, K.; Feng, Z.; Gilliland, G. L.; Iype, L.; Jain, S.; Fagan, P.; Marvin, J.; Padilla, D.; Ravichandran, V.; Schneider, B.; Thanki, N.; Weissig, H.; Westbrook, J. D.; Zardecki, C. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2002**, *58*, 899−907.
(58) Hayward, S.; Kitao, A.; Gō, N. *Protein Sci.* **1994**, *3*, 936−943.

DOI: 10.1021/acs.jctc.5b00724
J. Chem. Theory Comput. 2015, 11, 5503−5512