Molecular Simulation, Vol. 32, Nos. 10–11, 30 August–15 September 2006, 803–815

Annotated with a comment on page 811.

Simulating nanoscale functional motions of biomolecules

W. WRIGGERS†*, Z. ZHANG†, M. SHAH‡ and D. C. SORENSEN‡

†Laboratories for Biocomputing and Imaging, School of Health Information Sciences & Institute of Molecular Medicine, University of Texas Health Science Center at Houston, 7000 Fannin Street, Suite 600, Houston, TX 77030-5400, USA
‡Department of Computational and Applied Mathematics, Rice University, 6100 Main Street, Houston, TX 77005-1892, USA

(Received January 2006; in final form April 2006)

We are describing efficient dynamics simulation methods for the characterization of functional motion of biomolecules on the nanometer scale. Multivariate statistical methods are widely used to extract and enhance functional collective motions from molecular dynamics (MD) simulations. A dimension reduction in MD is often realized through a principal component analysis (PCA) or a singular value decomposition (SVD) of the trajectory. Normal mode analysis (NMA) is a related collective coordinate space approach, which involves the decomposition of the motion into vibration modes based on an elastic model. Using the myosin motor protein as an example we describe a hybrid technique termed amplified collective motions (ACM) that enhances sampling of conformational space through a combination of normal modes with atomic level MD. Unfortunately, the forced orthogonalization of modes in collective coordinate space leads to complex dependencies that are not necessarily consistent with the symmetry of biological macromolecules and assemblies. In many biological molecules, such as HIV-1 protease, reflective or rotational symmetries are present that are broken using standard orthogonal basis functions. We present a method to compute the plane of reflective symmetry or the axis of rotational symmetry from the trajectory frames. Moreover, we develop an SVD that best approximates the given trajectory while respecting the symmetry. Finally, we describe a local feature analysis (LFA) to construct a topographic representation of functional dynamics in terms of local features. The LFA representations are low-dimensional, and provide a reduced basis set for collective motions, but unlike global collective modes they are sparsely distributed and spatially localized. This yields a more reliable assignment of essential dynamics modes across different MD time windows.

Keywords: Principal component analysis; Normal modes; Singular value decomposition; Symmetry constraints; Local feature analysis; Molecular dynamics

1. Introduction: molecular dynamics and the sampling problem

One of the top ten challenges in computational biology is the prediction and engineering of function from structure of complex molecules [1]. This is not yet feasible on a routine basis, but novel simulation and mining technologies bring the prediction of function from structure within reach of biomedical researchers. The modeling of large nanometer scale functional motions will lead to a precise understanding of mutations and other biological variations, and the ability to design molecules for medical nanotechnology. We investigate the conformational dynamics of proteins and large-scale cellular machines [2] that is relevant for their function. Nanoscale biomolecules have been termed molecular machines due to their property to undergo conformational changes while transducing chemical energy into mechanical energy. A casual glance into today's journals reveals that concurrent structural biology is, to large extent, concerned with the quest to identify the moving parts of a biomolecular machine—its springs, shafts, levers, and axles.

The motion of large biomolecules is notoriously difficult to detect and to predict with conventional experimental observations alone. Molecular dynamics (MD), involving the numerical integration of Newton's equations of motion, is an important computational tool in the study of the functional dynamics [3,4]. Once nanoscale motions are generated or reproduced *in silico*, the functional "machine parts" can be visualized in computer graphics and structural fluctuations and transitions of biomolecules can be described at the atomic level.

One of the well-known and major limitations of MD in the application to large biomolecules is the shortness of achievable simulation times, typically of the order of tens to hundreds of nanoseconds, due to the femtosecond

*Corresponding author. Tel.: 713 500-3961. Fax: 713 500-3907. Email: wriggers@biomachina.org

Molecular Simulation
ISSN 0892-7022 print/ISSN 1029-0435 online © 2006 Taylor & Francis
http://www.tandf.co.uk/journals
DOI: 10.1080/08927020600771415

integration time step needed to time-resolve chemical bond vibrations. These times are much shorter than the time scales of many important biological processes, such as multi-domain motions and allosteric transitions. In the mid 1990s it became clear that correlations in low-frequency displacements are under sampled by nanosecond MD simulations, which prompted the question "how long is long enough?" [5]. An answer may be found in experimental studies that suggest that the relaxation times of correlations for multi-domain proteins are on the order of milliseconds or longer [6,7]. Therefore, the difficulty associated with their sampling limits the reliable prediction of such nanoscale functional motions using the traditional simulation techniques on much shorter time scales.

The significant advancements of recent years in the structural biology of large macromolecular assemblies of >10 nm length, such as cytoskeletal filaments, nuclear pores, the transcription machinery, and virus capsids [8], have exposed a gap between MD capabilities and desired biological relevance. This gap calls for novel computational methods that adequately represent the flexibility of nanoscale biomolecules. Three areas of advancements can be distinguished: (1) an increase in the MD integration time step size, without significantly affecting the dynamics [9–16], would bridge the time scale gap between MD and relevant functional dynamics; (2) a reduction in the complexity of the molecular models (coarse graining) would bridge the spatial scale gap [17–22] and reduce the number of degrees of freedom that need to be explored in nanoscale systems; and (3) a more sophisticated approach to boost sampling efficiency through statistical mining and enhancement of the dynamics [23–30], would enable the prediction of nanoscale motions from short MD simulations. In this review of recent methods developed and implemented in our laboratories, we focus mainly on the last topic, although this division is not exclusive: some recent hybrid methods are based on principles from more than one category.

The organization of this paper is as follows. Firstly, we review well-known collective coordinate space techniques that provide a dimension reduction via orthogonal basis functions. Subsequently, we describe the hybrid amplified collective motions (ACM) technique that enhances conformational sampling through a low-frequency boost of vibrational modes. We discuss limitations of collective coordinates with respect to symmetry and coherence of the predicted motion, before offering two solutions to this problem. The first solution is an adaptation of singular value decomposition (SVD) such that symmetric modes can be extracted from—and symmetry be enforced on—the MD trajectory. The second solution involves a departure from global collective coordinates, instead we propose the use of a local basis functions for functional dynamics. Finally, we provide concluding remarks on the parameterization and future applicability of the described algorithms.

2. Protein dynamics in collective coordinate space

Over the years many researchers strived to extract "essential" functional features from the short MD trajectories, with the hope to describe the motion in terms of a small number of variables, sometimes called collective coordinates or essential degrees of freedom [31–37]. One widely used statistical approach to such dimensionality reduction is principal component analysis (PCA) [38,39], also known as the Karhunen–Loeve expansion [40] in time series analysis. PCA is based in part on the related theory of SVD. After introducing SVD and PCA, we will discuss similarities with the harmonic description of protein dynamics in normal mode analysis (NMA). All of these methods have in common that they use an orthogonal set of global basis functions for characterizing the dynamics, with the aim to truncate a series expansion for a reduced, low-dimensional approximation of the dynamics.

2.1 Singular value decomposition

Given a trajectory $\mathcal{S} = \{\mathbf{x}(t),\, t \geq 0\}$, we assume that the overall translational and rotational motion of the molecule is eliminated from the trajectory as is customary in MD, and that the mean of internal motion has been subtracted from the coordinates, i.e. the trajectory $\mathcal{S}$ is centered at the origin of $\mathbb{R}^{3N}$, where N is the number of atoms under consideration. Often the full set of atoms is used, but assuming a protein structure, one might consider only the C_α atoms, since Amadei *et al.* demonstrated that the larger amplitude "essential" modes are robust under such a coarse graining [33,41].

For a system of N considered atoms, the Gramian (also named covariance matrix or second moment matrix)

$$\mathbf{C} = \int_0^{T_f} \mathbf{x}(\tau)\mathbf{x}(\tau)^{\mathrm{T}}\,\mathrm{d}\tau \tag{1}$$

is a $3N \times 3N$ symmetric positive semi-definite matrix. The eigensystem of $\mathbf{C}$

$$\mathbf{C} = \mathbf{U}\mathbf{S}^2\mathbf{U}^{\mathrm{T}} \tag{2}$$

provides an orthogonal basis via the columns of $\mathbf{U}$ and in this basis we have the representation

$$\mathbf{x}(t) = \mathbf{U}\mathbf{S}\mathbf{v}(t)$$

with the components of $\mathbf{v}(t)$ being mutually orthogonal square integrable functions on $(0,\, T_f)$. If the diagonal elements of $\mathbf{S}$ decay rapidly (assuming they are in decreasing order) then a reduced basis representation of the trajectory may be obtained by discarding the trailing terms and considering the approximation $\mathbf{x}_n = \mathbf{U}_n\mathbf{S}_n\mathbf{v}_n(t)$, where the subscript n denotes the leading n columns and/or components ($n \ll 3N$).

This continuous case is approximated using snapshots consisting of values $\mathbf{x}(t_j)$ of the trajectory at discrete time

points t_j and forming the $3N \times m$ matrix

$$\mathbf{X} = [\mathbf{x}(t_1), \mathbf{x}(t_2), \ldots, \mathbf{x}(t_m)].$$

The SVD of $\mathbf{X}$ provides

$$\mathbf{X} = \mathbf{U}\mathbf{S}\mathbf{V}^{\mathrm{T}} \approx \mathbf{U}_n\mathbf{S}_n\mathbf{V}_n^{\mathrm{T}}$$

where

$$\mathbf{U}^{\mathrm{T}}\mathbf{U} = \mathbf{V}^{\mathrm{T}}\mathbf{V} = \mathbf{I}, \mathbf{S} = \mathrm{diag}(\sigma_1, \sigma_2, \ldots, \sigma_{3N}),$$

$\mathbf{I}$ denotes the identity matrix, and the diagonal elements are ordered $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_{3N}$. The left singular vectors (columns of $\mathbf{U}_n$) of a truncated SVD then provide the reduced basis. It can be shown that $\mathbf{U}$ and $\mathbf{S}$, resulting from SVD, are identical to the eigensystem of $\mathbf{C}$ in equation (2).

The approximation of the idealized $\mathbf{C}$ to the discrete time steps

$$\mathbf{C} \approx \frac{1}{m}\mathbf{X}\mathbf{X}^{\mathrm{T}} \equiv \langle \mathbf{x}\mathbf{x}^{\mathrm{T}} \rangle$$

is provided by a quadrature rule. Here and in the following $\langle\rangle$ denotes an average over the time frames.

2.2 Principal component analysis

In PCA, we take advantage of the orthogonality of the columns of $\mathbf{U}$, $\mathbf{u}_r (r = 1, \ldots, 3N)$, that form a basis for the internal motion of the molecule. The basis functions are often termed principal modes if resulting from the eigensystem of $\mathbf{C}$, or left-singular vectors if resulting from SVD.

This statistical method was introduced in a related form discussed below to the biomolecular simulation community by McCammon, Karplus and their coworkers [42,43] in the 1980s, initially under the name quasi-harmonic analysis. Later, in the early 1990s the idea witnessed a renaissance under the name essential dynamics and has since enjoyed the increasing enthusiasm of a large number of investigators [33,34] who successfully applied it to sample the conformational space [36,37] and to investigate the physical nature of protein dynamics.

Due to orthogonality of the principal modes the components of the coordinates $\mathbf{x}$ can be reconstructed from the modes:

$$x_i = \sum_{r=1}^{3N} A_r u_r(i) \quad \text{with}$$

$$A_r = \sum_{i=1}^{3N} u_r(i) x_i \equiv \sum_{i=1}^{3N} K_r(i) x_i, \tag{3}$$

where A_r is the so-called output of the representation, i.e. the projection of atomic fluctuations onto the principal mode $\mathbf{u}_r$. PCA outputs are decorrelated in the sense that $\langle A_r A_s \rangle = \sigma_r^2 \delta_{rs}$. $K_r(i)$ is the so-called kernel of the PCA representation, in the case of PCA $K_r(i) = u_r(i)$. As explained above, the eigenvalues σ_r^2 are ordered in a decreasing sequence, and we assume that a small number n ($n \ll 3N$) of modes are sufficient to describe the dominant dynamics. This means we truncate the expansion (equation (3)) early and define the (approximate) reconstructed coordinates:

$$x_i^{\mathrm{rec}} = \sum_{r=1}^{n} A_r u_r(i). \tag{4}$$

2.3 Quasi-harmonic and normal mode analysis

The original introduction of PCA to molecular modeling by McCammon, Karplus, and their coworkers [42,43] relied on a quasi-harmonic approximation of the potential energy that would reproduce the observed second moments of the atomic displacements $\mathbf{C}$ (equation (1)). The quasi-harmonic force constant matrix, $\mathbf{F}$, can be written

$$\mathbf{F} = k_{\mathrm{B}} T \mathbf{C}^{-1}, \tag{5}$$

where k_{B} is the Boltzmann constant and T is the temperature. In the quasi-harmonic approximation the potential energy $W(\mathbf{x})$ of the system varies quadratically about a given equilibrium conformation $\mathbf{x}_0$. Since the gradient of the potential vanishes at $\mathbf{x}_0$, the Taylor series expansion of W about $\mathbf{x}_0$ is given by:

$$W(\mathbf{x}) = W(\mathbf{x}_0) + \frac{1}{2}(\mathbf{x} - \mathbf{x}_0)^{\mathrm{T}}\mathbf{F}(\mathbf{x} - \mathbf{x}_0) + \cdots \tag{6}$$

If third and higher order derivatives are ignored the dynamics of the system can be described in terms of the eigenvectors (normal modes) Ω_r and frequencies ω_r ($r = 1, \ldots, 3N$). In Cartesian coordinates the Ω_i, ω_i satisfy the eigensystem [44]

$$\mathbf{M}^{-1/2}\mathbf{F}\mathbf{M}^{-1/2}\Omega_r = \omega_r^2 \Omega_r$$
$$\Omega_r \Omega_s = \delta_{rs} \tag{7}$$

where $\mathbf{M}$ is a diagonal matrix of atomic masses m_i. In Cartesian coordinates solving equation (7) for N atoms involves again a numerical diagonalization of a $3N \times 3N$ matrix. If $m_i = \mathrm{const}$ (as in coarse-grained C_α based representations), normal modes and principal modes from PCA are identical. The normal modes Ω_r are ordered here by increasing frequencies $\omega_1 \leq \omega_2 \leq \cdots \leq \omega_{3N}$. For a given set of initial amplitudes α_r and phases ϕ_r at time $t = 0$, the time evolution in the normal mode formulation is then [44]

$$x_i(t) = x_i(0) + \sqrt{2}\sum_{r=1}^{3N} \Omega_r(i) m_i^{-1/2} \alpha_r \cos(\omega_r t + \phi_r). \tag{8}$$

It can be shown using statistical mechanics that the amplitudes at thermal equilibrium are inversely proportional to frequency [45]. Thus, the largest 1% of the modes contributes up to 90% of the atomic fluctuations. In practical situations it is therefore again sufficient to extract only $n \ll 3N$ lowest frequency modes.

While there is general agreement about the appeal of PCA and quasi-harmonic analysis for the prediction of functionally relevant modes, it became clear in the 1990s that such theories suffer from the MD sampling problem. For large systems of interest the second moments **C** may not converge on the short time scales accessible to MD. Also, García *et al.* demonstrated that for large systems the distribution of conformations often becomes multi-modal [35] (as suggested also by Go's "jumping-among-minima model" [46]). For such multi-modal distributions the Boltzmann inversion (equation (5)) and harmonic approximation (equation (6)) break down. Principal modes derived from different MD time windows are therefore no longer consistent [47]. Therefore, for the large systems of interest here, PCA and quasi-harmonic modes from short MD trajectories are intrinsically unreliable.

An alternative approach to the (unreliable) sampling of quasi-harmonic modes from MD can be obtained if one directly postulates a harmonic approximation of the potential energy in the vicinity of an equilibrium position $\mathbf{x}_0$. In what has become known as conventional NMA or elastic network theory, the statistically sampled **F** is replaced by a Hessian matrix of second derivatives of a potential energy model (e.g. the standard MD potential energy evaluated at $\mathbf{x}_0$) [44,48]. This idea is rooted in the observation that biomolecules often behave more than one might expect as if the energy surface were harmonic, even though, as we know, the potential really contains many local minima and anharmonic contributions. Conventional NMA, by means of the postulated Hessian, returns $3N - 6$ usable normal modes that describe the internal dynamics (since rigid-body motions are not removed from the Hessian, the first six frequencies are zero and the corresponding rigid-body modes $r = 1, \ldots, 6$ are usually ignored). This way, animations of large-scale vibrational modes can be produced (via equation (8)) that often agree surprisingly well with experimentally observed motions of biomolecules [49]. Although, conventional NMA is based on an entirely heuristic elasto-mechanic model, it has become well established in biophysical simulation and refinement, mainly due to the ease by which appealing animations of large-scale motions are generated from static biomolecular structures. Although both are rooted in normal mode analysis, in the following "NMA" and "normal mode" refers to the widely used, strictly harmonic convention, whereas "PCA" and "principal mode" refers to the quasi-harmonic analysis based on MD.

2.4 Comparison of collective coordinate methods

Figure 1 provides a comparison of the first principal mode from an MD simulation of myosin with the first non-trivial normal mode. The initial structure for all myosin simulations in this work was taken from the supplementary structure "motor_domain.pdb" published by Holmes *et al.* [50]. NMA was performed with a course-grained model as described in [30]. MD was performed using the GROMOS96 [51] simulation package with an united-atom parameter set 43A1 (Z.Z. and W.W., to be published). The SPC water model was used to describe the solvent molecules [52]. The system consists of myosin (1099 residues and 11,216 atoms), 32 Na^+ ions, and 57,650 water molecules, leading to a total size of 184,198 atoms. After energy minimization and 100 ps positional-restraint equilibration, a 1 ns production simulation was performed. The three groups (protein, ions and solvent) were coupled separately to a temperature bath of reference temperature 300 K (relaxation time 0.1 ps) [53].

The normal mode (figure 1(b)) corresponds to the well-known lever arm motion of the myosin motor protein [54] and is slowly varying across the molecule, whereas the principal mode (figure 1(b)) is noisy and less coherent due to under sampling of the large scale variability of myosin in the MD trajectory. It is apparent that for a large system of 14 nm size such as myosin S1, a standard nanosecond MD simulation is not long enough to lead to a convergence of principal and normal modes, although such a convergence was proposed in the literature for smaller systems [31].

3. Enhanced sampling by hybrid strategies

A modeler who wishes to employ collective coordinate methods usually faces a tradeoff: If simulations of structural rearrangements are desired that involve the forming and breaking of contacts between secondary structure elements or domains, one is limited to MD techniques with their known under sampling of large conformational changes. If on the other hand the simulation of large scale motion is desired, one may wish to employ NMA, but the elastic model used in NMA dues not provide for local relaxation of the stereochemistry, or for forming or breaking of contacts in the structure. These limitations have prompted Zhang *et al.* to develop a hybrid method, termed ACM that combines the advantages of both NMA and MD while eliminating some of their limitations.

In [30] the authors used the normal modes obtained with a coarse-grained elastic network model to guide the atomic-level MD simulations. Based on NMA, collective modes can be calculated according to the current snapshot **X** in the conformational space of the protein. The authors then project the velocities of all atoms in the system on a subspace of $n \ll 3N$ normal modes. In the MD simulation protocol, the velocities along the slowest few modes are coupled to a higher temperature by means of Berendsen's weak coupling method [53] to amplify the collective motions. Velocities orthogonal to this subspace are coupled to standard room temperature. This selective coupling to two different temperatures has the advantage that only a very small number of the degrees of freedom are experiencing the elevated temperature that would normally denature a protein, whereas the large remainder of orthogonal degrees of freedom dissipates the thermal energy efficiently, leading to a stable dynamics that

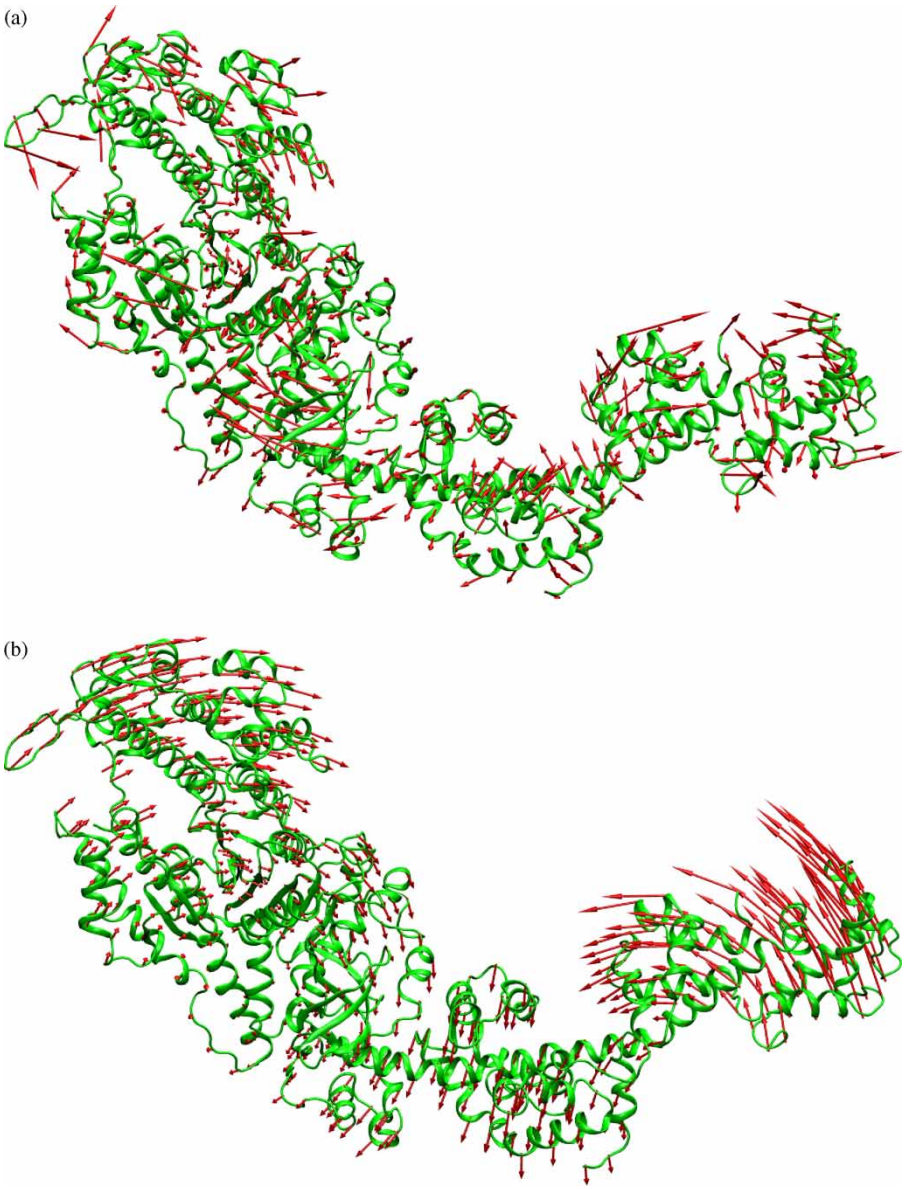

Figure 1. Comparison of PCA and NMA of the S1 motor domain of myosin II. (a) The first ($r = 1$) principal mode (see text) extracted from the MD simulation. (b) The first non-trivial ($r = 7$) normal mode. All figures in this paper are available in colour online and all were created with the visualization program VMD [65].

is guided towards the low-frequency modes of interest. In [30] this low-boost sampling technique was applied to two test systems. One was a 15 amino acid S-peptide analog, where it was possible to perform reversible folding. MD simulations typically denature biomolecules irreversibly, but ACM was sufficiently powerful to refold the denatured peptides back into their native structure. The other system was bacteriophage T4 lysozyme. Much more extensive domain motions between the N-terminal and C-terminal domain of T4 lysozyme were observed in the ACM simulation compared to a conventional simulation. The only disadvantage of ACM is the non-equilibrium character of the dynamics due to the energy flow from slow to fast modes. This makes it difficult to estimate a Boltzmann distribution and potential of mean force as is needed for free energy calculations. However, for conformational sampling it appears the technique appears to be quite powerful.

In this work we have modified the GROMOS96 package [51] to implement ACM as described. The ACM simulation protocol of myosin was identical to that of the standard simulation described above, except for temperature coupling. The first three collective modes were coupled to a higher temperature 800 K with a relaxation time of 0.006 ps, and the other degrees of freedom were coupled to the room temperature 300 K (relaxation time 0.1 ps). The collective modes were updated every 250 time steps according to the changing configuration of myosin.

Figure (2) provides a comparison of the conformational sampling achieved with standard MD and ACM. It is well known that the lever arm of myosin is flexible during

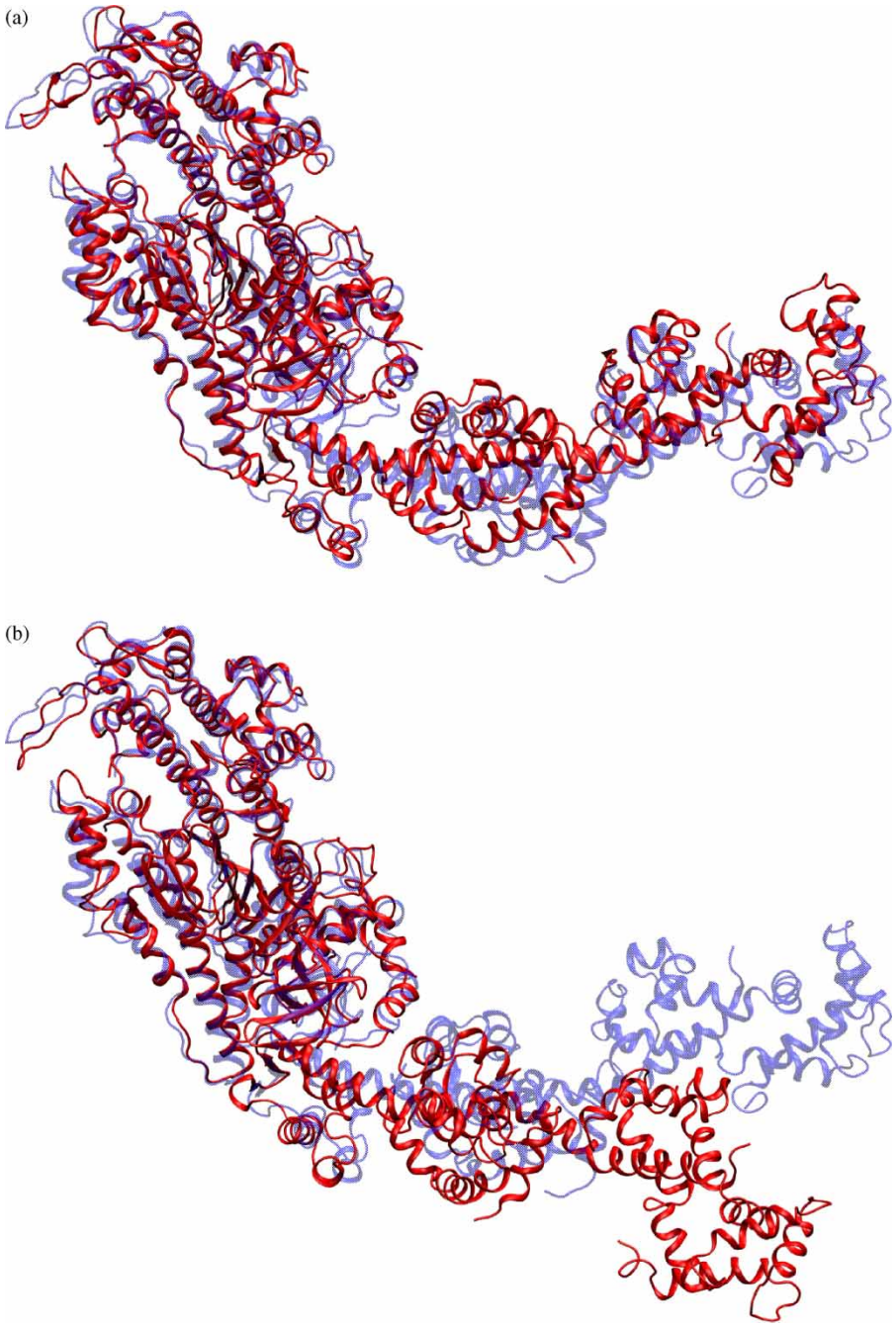

Figure 2. Effect of ACM on myosin dynamics. (a) Standard MD. Blue: the initial structure, and red: the snapshot $t = 992$ ps that exhibits the largest rms deviation from the initial structure among all trajectory frames. (b) ACM simulation. Blue: the initial structure, and red: the snapshot $t = 651$ ps that exhibits the largest rms deviation. The structures in (a) and (b) were least-squares fitted by the C_α atoms of res. 1–607 (2MYS [66] numbering).

muscle contraction and coupled mechanically to the distant active site. However, in standard MD the large system remains close to its initial configuration. The lever arm rotates only by about 6°, which corresponds to a maximum shift of 1 nm (figure 2(a)). However, in the ACM simulation, the lever rotates about 31° relative to the head, which leads to an overall displacement of 5 nm. This variability is in excellent agreement with experimental results. Electron microscopy has shown a 32° rotation [55], whereas a nanomanipulation of single myosin heads attached to actin filaments showed a displacement of 5.3 nm per power stroke [56]. Although one would not necessarily expect ACM snapshots of a chemically inactive myosin to resemble the chemically active motor, it is clear from figure 2 that ACM is able to overcome the sampling problem of MD for large systems while providing for a relaxation of tertiary structure that can be interpreted in terms of an allosteric mechanism.

4. Limitations of orthogonal basis functions

Collective coordinates methods of Section 2 either estimate (in case of SVD or NMA) or postulate (in case

of NMA) a coherence of motion across large distances in biomolecules. Since the sampling of such inter-domain coherence is out of reach for short MD simulations (figure 1), PCA and SVD risk to overestimate such a coherence in an under sampling situation. Also, normal modes may suffer from inaccuracies since they are based on a heuristic elastic model that is not rooted in first principles of biomolecular interactions. The major advantage of collective coordinate methods is their suitability for dimensionality reduction through truncation of a series expansion (cf. equation (4)). However, the inaccuracies in the modes (due to under sampling or empirical modeling) may conspire to obscure the relevance of individual modes, such that it is not clear *a priori* which modes or which linear combination of modes (in a given low-dimensional subspace) are functionally relevant [57].

The global extent of individual collective modes is problematic not only because of the limited reliability of individual modes, but also because of their forced orthogonalization. Since the rth mode is always forced to be orthogonal to the first r-1 modes, complex causal dependencies arise such that a particular mode r is actually dependent on modes of a lower index. In the case of PCA it was shown that even fast ($r \gg 1$) modes, whose relaxation time is well within the MD sampling window, cannot be recovered by PCA due to their dependence on the slower, under sampled modes [47]. The same causal dependencies due to orthogonality can also be expected for NMA, since any noise in the assignment of low-frequency modes will cascade throughout the entire orthogonal system of basis functions.

The forced orthogonalization of modes also has the undesirable effect of breaking the symmetry of large-scale macromolecular assemblies. For example, a three-fold symmetric system such as tricorn protease [58] should exhibit a symmetry related representation (for each 120° rotation). Instead an orthogonal transform fixes by numeric chance one of the three possible solutions and forces all subsequent modes to be orthogonal, thereby breaking the symmetry.

Due to the apparent limitations of orthogonal collective coordinates we were seeking alternative statistical theories that do not suffer from the orthogonalization problem while obeying the symmetry of a given system. In the following two section we provide an abridged overview of two such theories that have been introduced very recently [59,60]. The first strategy based on SVD still uses global basis functions but provides approximate modes of motion that best describe the symmetric movements of the protein. The second approach is more radical, instead of global orthogonal modes new local basis functions are constructed. These local feature representations are still low-dimensional and provide a reduced basis set for collective motions, but they are sparsely distributed and spatially localized. The theory is augmented by new application results on the myosin motor.

5. SVD approximation with symmetry constraints

In the following we concentrate on detecting and enforcing two types of symmetry within the SVD formalism: rotational and reflective. The computational schemes for calculating the best symmetric approximation of a given trajectory in $\mathbb{R}^{3N}$ space involves two stages for each case. For reflective symmetry, the first stage is to obtain the normal $\mathbf{w}$ to an approximate plane of reflective symmetry, where the normal is defined to be the unit vector perpendicular to a hyperplane $\mathcal{H}$ for which the given set of trajectory frames $\mathcal{S}$ can be split into two mirror sets. For rotational symmetry, we first determine an approximate axis of rotational symmetry $\mathbf{q}$ about which the given set can be rotated in steps of $2\pi/k$. In the second stage, we find the best approximation to the given trajectory that has the appropriate symmetries enforced. Since simulation trajectories are noisy and do not strictly obey the expected symmetry, we need to construct a normal vector or axis of rotation that diminishes the effects of outliers. This is accomplished by an iterative re-weighting scheme that minimizes deviation from symmetry in a weighted norm. We also provide a means to compute just the dominant portion (leading n terms) of the SVD that is well suited to large scale computation. This computation only requires matrix–vector products involving the trajectory frame set represented as a matrix. The ARPACK software [61] can be used in this large-scale case. As demonstrated on a model system, the computation is no more expensive than constructing the leading terms of the SVD of the full set of points without the symmetry constraint. Complete details concerning this methodology and its implementation may be found in a technical report [59].

5.1 Stage 1: reflective symmetry case

Recall that a hyperplane $\mathcal{H}$ is specified by a constant γ and a vector $\mathbf{w}$ via $\mathcal{H} := \{\mathbf{x} : \gamma + \mathbf{w}^{\mathrm{T}}\mathbf{x} = 0\}$. The vector $\mathbf{w}$ is called the normal to the plane. The symmetry relation is described by the orthogonal transformation $\mathbf{I}_3 - 2\mathbf{w}\mathbf{w}^{\mathrm{T}}$ which is known as an elementary reflector, where $\mathbf{I}_3$ is the 3×3 identity matrix. A set of points $\mathcal{P} \in \mathbb{R}^3$ is reflectively symmetric with respect to a hyperplane $\mathcal{H}$ with unit normal $\mathbf{w}$ if and only if

$$\mathcal{P} = (\mathbf{I}_3 - 2\mathbf{w}\mathbf{w}^{\mathrm{T}})\mathcal{P}.$$

If our trajectory $\mathcal{S}$ is reflectively symmetric about $\mathcal{H}$, we can arrange the points of $\mathcal{S}$ into two sets represented as matrices $\mathbf{X}_0$ and $\mathbf{X}_1$ such that

$$\mathbf{X}_0 = \mathbf{W}\mathbf{X}_1,$$

where $\mathbf{W}$ is an order $3(N/2)$ block diagonal orthogonal matrix with the 3×3 orthogonal matrix $\hat{\mathbf{W}} = \mathbf{I}_3 - 2\mathbf{w}\mathbf{w}^{\mathrm{T}}$ in each diagonal block.

In general the given set of trajectory frames $\mathcal{S}$ is not exactly symmetric with respect to any particular plane due

to noise. However, we may think of calculating a normal **w** that does the best possible job of specifying a plane that separates $\mathcal{S}$ into two sets $\mathbf{X}_0$ and $\mathbf{X}_1$, that are "nearly" symmetric with respect to the plane. We assume that a partitioning of $\mathcal{S}$ into $\mathbf{X}_0$ and $\mathbf{X}_1$, is given such that the columns of the two matrices are correctly paired.

To find the normal **w** to the plane of symmetry, we rearrange the data into two related matrices $\mathcal{X}_0$ and $\mathcal{X}_1$ with

$$\mathcal{X}_j = \Big[\mathbf{X}_j(1:3, 1:m), \mathbf{X}_j(4:7, 1:m), \ldots, \mathbf{X}_j \left(\frac{3N}{2} - 2 : \frac{3N}{2}, 1:m\right)\Big].$$

Here, each $\mathbf{X}_j(i{:}i+2, 1{:}m)$ is a $3 \times m$ matrix consisting of the displacement coordinates of a particular atom throughout the trajectory. The specification of **w** may be expressed as an optimization problem

$$\min_{\|\mathrm{w}\|=1} \left\{ \|(\mathcal{X}_0 - \hat{\mathbf{W}}\mathcal{X}_1)\mathbf{D}\|_{\mathrm{F}} : \hat{\mathbf{W}} = \mathbf{I}_3 - 2\mathbf{w}\mathbf{w}^{\mathrm{T}} \right\}, \qquad (9)$$

where **D** is a diagonal weighting matrix and $\|.\|_{\mathrm{F}}$ denotes the Frobenius norm [59].

The weighting **D** is introduced to provide a means to de-emphasize anomalies and outliers in the supposed symmetry relation. If **D** is given, then the minimization can be solved. It turns out that the solution **w** to the minimization problem (equation (9)) is the unit eigenvector corresponding to the smallest eigenvalue of the symmetric indefinite matrix

$$\mathbf{M} = \mathcal{X}_0\mathbf{D}^2\mathcal{X}_1^{\mathrm{T}} + \mathcal{X}_1\mathbf{D}^2\mathcal{X}_0^{\mathrm{T}}.$$

We have devised in [59] an iterative re-weighting scheme that constructs an optimal weighting **D** and hence specifies the plane that best describes the symmetry condition.

5.2 Stage 1: rotational symmetry case

A set of points $\mathcal{P} \in \mathbb{R}^3 \cap \mathbf{q}^{\perp}$ is said to be k-fold rotationally symmetric about an axis $\mathbf{q} \in \mathcal{R}^3$ if there exist an orthogonal transformation $\hat{\mathbf{R}}(\mathbf{q})$ such that for every point $\mathbf{p} \in \mathcal{P}$, there exists $k-1$ distinct points $\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_{k-1} \in \mathcal{P}$, such that $\hat{\mathbf{R}}(\mathbf{q})^l\mathbf{p} = \mathbf{p}_l$ for $l = 1, 2, \ldots, k-1$. We call **q** the rotational axis of symmetry and $\hat{\mathbf{R}}(\mathbf{q})$ the rotation matrix. It can be shown that a set $\mathcal{S}$ is k-fold rotationally symmetric with respect to a rotational axis q if and only if for $l = 1, 2, \ldots, k-1$

$$\mathcal{P} = \hat{\mathbf{R}}(\mathbf{q})^l\mathcal{P} = (\mathbf{I}_3 - \mathbf{Q}\mathbf{G}_k\mathbf{Q}^{\mathrm{T}})^l\mathcal{P}.$$

where $[\mathbf{q}, \mathbf{Q}] \in \mathbb{R}^{3\times3}$ is an orthogonal matrix, and $\mathbf{I}_2 - \mathbf{G}_k \in \mathbb{R}^{2\times2}$ rotates any point $\mathbf{p} \in \mathbb{R}^3$ by an angle $\theta = 2\pi/k$ about the origin.

In the presence of noise we need to calculate a rotational axis **q** that best fits the data. When there is k-fold rotational symmetry present in the trajectory $\mathcal{S}$, we may assume a partitioning of $\mathcal{S}$ into $\mathbf{X}_0, \mathbf{X}_1, \ldots, \mathbf{X}_{k-1}$ such that the columns of the matrices are correctly paired with respect to rotation.

To express the optimality condition that will specify **q**, it is again convenient to reorganize the data into matrices $\mathcal{X}_0, \mathcal{X}_1, \ldots, \mathcal{X}_{k-1}$ as was done in the reflective case. The optimization problem

$$\min_{\|\mathbf{q}\|=1} \left\{ \left\| \mathbf{q}^{\mathrm{T}} \left[(k-1)\mathcal{X}_0 - \sum_{l=1}^{k-1} \mathcal{X}_l \right] \right\|_{\mathrm{F}} \right\} \qquad (10)$$

will correctly identify the rotational axis of symmetry **q** in the case of exact symmetry and will determine an optimal approximation in the presence of noise.

The solution **q** to the minimization problem (equation (10)) is the unit eigenvector corresponding to the smallest eigenvalue of $\mathbf{M}\mathbf{M}^{\mathrm{T}}$, where

$$\mathbf{M} = (k-1)\mathcal{X}_0 - \sum_{l=1}^{k-1} \mathcal{X}_1. \qquad (11)$$

Like in the case of reflective symmetry, we can introduce into the optimization a weighting scheme that minimizes the influence of outliers in the supposed rotational symmetry relation:

$$\min_{\|\mathbf{q}\|=1} \left\{ \left\| \mathbf{q}^{\mathrm{T}} \left[(k-1)\mathcal{X}_0 - \sum_{l=1}^{k-1} \mathcal{X}_l \right] \mathbf{D} \right\|_{\mathrm{F}} \right\} \qquad (12)$$

where **D** is a diagonal weighting matrix. The solution to equation (12) is the unit eigenvector **q** corresponding to the smallest eigenvalue of $\mathbf{M}\mathbf{D}^2\mathbf{M}^{\mathrm{T}}$, where **M** is defined as in equation (11). Again, there is an iterative re-weighting scheme to determine the optimal value of **q** (see [59] for details). With exact symmetry, the trajectory $\mathcal{S}$ should satisfy

$$\mathbf{X}_{j+1} = \mathbf{R}\mathbf{X}_j, \quad \text{for} \quad j = 0, 1, \ldots, k-1,$$

where **R** is an order $3(N/k)$ block diagonal orthogonal matrix with diagonal blocks consisting of the 3 by 3 matrix $\hat{\mathbf{R}} = \mathbf{I}_3 - \mathbf{Q}\mathbf{G}_k\mathbf{Q}^{\mathrm{T}}$.

5.3 Stage 2: symmetric approximation of the trajectory

To find the best reflective or rotational symmetric approximation to a set of trajectory frames, we can take advantage of the following result [59]. For reflective symmetry $\mathbf{R} = \mathbf{W}$ and $\mathbf{W}^2 = \mathbf{I}$, and in the case of rotational symmetry $\mathbf{R} = \mathbf{R}(\mathbf{q})$ and $\mathbf{R}(\mathbf{q})^k = \mathbf{I}$.

If

$$\mathbf{R}^{k-l}\mathbf{X}_l = \mathbf{X}_0 + \mathbf{E}_l,$$

where $\mathbf{E}$ represents the deviation from ideal symmetry and $\mathbf{R}^k = \mathbf{I}$, then

$$\min_{\hat{\mathbf{X}}_{h+1}=\mathbf{R}\hat{\mathbf{X}}_h} \left\| \begin{pmatrix} \mathbf{X}_0 \\ \vdots \\ \mathbf{X}_{k-1} \end{pmatrix} - \begin{pmatrix} \hat{\mathbf{X}}_0 \\ \vdots \\ \hat{\mathbf{X}}_{k-1} \end{pmatrix} \right\|^2 = \frac{1}{k} \sum_{l=0}^{k-1} \sum_{h=l+1}^{k-1} \|\mathbf{E}_h - \mathbf{R}^{h-l}\mathbf{E}_l\|_{\mathrm{F}}^2,$$

and the SVD of this optimal solution (denoted by $\hat{}$)

$$\begin{pmatrix} \hat{\mathbf{X}}_0 \\ \vdots \\ \hat{\mathbf{X}}_{k-1} \end{pmatrix} = \mathrm{USV}^{\mathrm{T}},$$

satisfies

$$\mathbf{U} = \frac{1}{\sqrt{k}} \begin{pmatrix} \mathbf{U}_0 \\ \vdots \\ \mathbf{U}_{k-1} \end{pmatrix}, \quad \mathbf{S} = \sqrt{k}\mathbf{S}_0, \quad \mathbf{V} = \mathbf{V}_0,$$

where

$$\mathbf{U}_l = \mathbf{R}^l\mathbf{U}_0, \quad l = 0, 1, 2, \ldots, k-1,$$

and

$$\mathbf{U}_0\mathbf{S}_0\mathbf{V}_0^{\mathrm{T}} = \frac{1}{k}(\mathbf{X}_0 + \mathbf{R}^{k-1}\mathbf{X}_1 + \mathbf{R}^{k-2}\mathbf{X}_2 + \cdots + \mathbf{R}\mathbf{X}_{k-1}).$$

5.4 SVD approximation results

The algorithmic structure for both the reflective and rotational SVD approximation is the same. To summarize, it consists of the two stages

1. Determine the normal $\mathbf{w}$ or the axis $\mathbf{q}$ for reflective or rotational symmetry, respectively.
2. Compute the standard SVD

$$\mathbf{U}_0\mathbf{S}_0\mathbf{V}_0^{\mathrm{T}} = \frac{1}{k}(\mathbf{X}_0 + \mathbf{R}^{k-1}\mathbf{X}_1 + \mathbf{R}^{k-2}\mathbf{X}_2 + \cdots + \mathbf{R}\mathbf{X}_{k-1})$$

where $\mathbf{R}$ is a reflector determined by $\mathbf{w}$ or a rotation about the axis determined by $\mathbf{q}$.

We seek the dominant (largest) singular values. This can be done in a straightforward manner using the ARPACK software on a serial computer or P_ARPACK on a parallel system. Only the leading n terms (singular values) are required. One may either specify n or utilize a restarting scheme to adjust n until $\sigma_n \geq \mathrm{tol}{*}\sigma_1 > \sigma_{n+1}$. The important computational point is that only matrix–vector products of the form

$$\mathbf{u} = \frac{1}{k}(\mathbf{X}_0 + \mathbf{R}^{k-1}\mathbf{X}_1 + \mathbf{R}^{k-2}\mathbf{X}_2 + \cdots + \mathbf{R}\mathbf{X}_{k-1})\mathbf{v}$$

need to be computed. This is essentially the same work one would require to compute the corresponding standard SVD of $\mathbf{X}$ without the symmetry constraint.

This analysis was carried out using P_ARPACK on a Linux cluster with six dual-processor nodes consisting of 1600 MHz AMD Athlon processors with 1 GB RAM per node and a 1 GB/s Ethernet connection. The method was applied to compute the leading modes for HIV-1 protease [62]. The system consists of 3120 atoms and hence there are 9360 degrees of freedom in the full-atom representation. The MD trajectory consisted of 10,000 time steps [59].

These computations were done for both reflective and rotational symmetry with essentially the same computational time needed as in the standard SVD case. For $n = 50$ singular values, the symmetry enforced SVD took 312 s, while the regular SVD took 390 s. The use of P_ARPACK to compute just the dominant n terms was essential for the high efficiency when dealing with the large structure.

Figure 3 shows a snapshot of HIV-1 protease comparing the SVD outputs from the first 10 rotationally symmetric modes (blue) with a standard SVD outputs using the first 10 regular modes of motion (red). HIV-1 protease has a two-fold rotational symmetry and this aspect is preserved in the symmetrized SVD while providing a good approximations to the standard trajectory.

6. Local feature analysis

The PCA and SVD representations above offer a reduced dimensionality, however, they are non-local. By this we mean that the kernel functions $K_r(i)$ (equation (3)) extend over the entire range of i (the $3N$ degrees of freedom of the biomolecule), but nearby values in the r index have no relationship among each other. In the following we recast the expansion into a new representation that obeys locality, i.e. the kernel functions are not labeled by the principal mode index r, but by the index of the degrees of freedom, i.

6.1 Abridged LFA theory

In [60] we derived the theory of local feature analysis (LFA) for biomolecular dynamics that can be formulated in a compact form as follows. If we define a family of matrices

$$\mathbf{K}^{(m)} = \mathbf{US}^{-m}\mathbf{U}^{\mathrm{T}},$$

strictly only for $m = -2$ or $n = 3N$, in other cases the summation on the right is truncated at n (see [60]).

the cases $m = 1, 0, -1, -2$ are of specific importance in LFA:

- $\mathbf{K}^{(1)}$ is the LFA kernel which satisfies locality: Similar to the PCA outputs A_r in equation (3), we can define local outputs $O(i)$

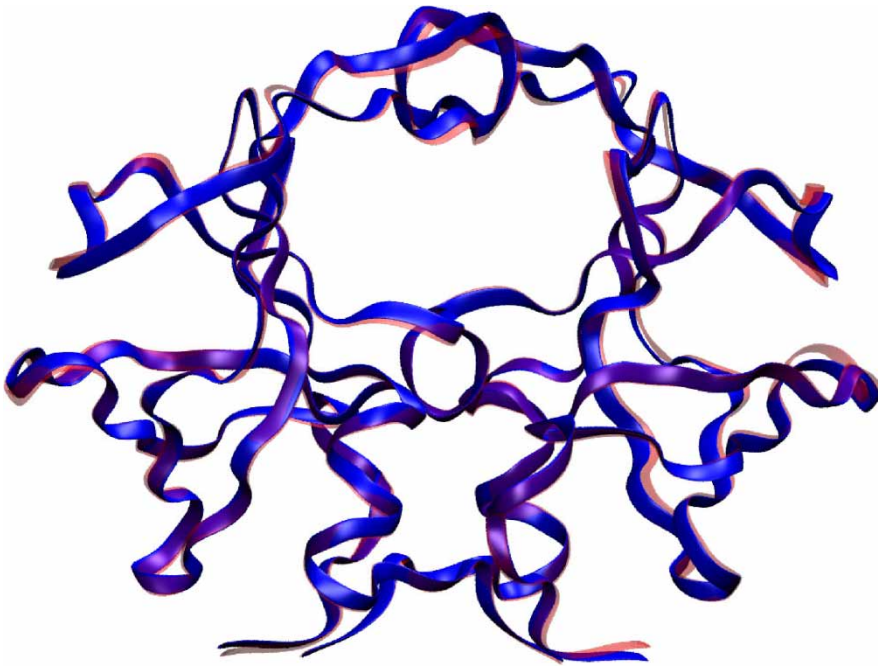

Figure 3. Comparison of standard SVD (red transparent ribbon) and symmetry enforced SVD (blue solid ribbon in online version) applied to a trajectory of HIV-1 protease. Shown is the projection of one trajectory frame on the subspace of the first 10 principal modes returned by either method (see text).

$$O(i) \equiv \sum_{j=1}^{3N} \mathbf{K}^{(1)}(i,j)\mathbf{x}_j,$$

but here, using the local kernel, O depends on i and not on r.

- $\mathbf{K}^{(0)}$ is the residual output correlation. Since the $3N$ outputs $O(i)$ are derived from only $n \ll 3N$ linearly independent principal modes, the LFA outputs are not fully decorrelated, instead one can show that

$$\langle O(i)O(j)\rangle = \mathbf{K}^{(0)}(i,j).$$

The LFA outputs become completely decorrelated ($\mathbf{K}^{(0)}(i,j) \to \delta(i,j)$) only in the limit $n \to 3N$.

- $\mathbf{K}^{(-1)}$ is the reconstructor (inverse LFA kernel):

$$\mathbf{x}_i^{\mathrm{rec}} \equiv \sum_{j=1}^{3N} \mathbf{K}^{(-1)}(i,j)O(j).$$

- $\mathbf{K}^{(-2)} \equiv \mathbf{C}$ is the covariance matrix (equation (1)), which follows trivially from equation (2).

The matrices $\mathbf{K}^{(1)}$ and $\mathbf{K}^{(0)}$ are central to LFA. $\mathbf{K}^{(1)}$ is by definition the projection operator onto a local feature. The resulting projections (or outputs) $O(i)$ are dimensionless and in the limit $n \to 3N$ become orthogonal, as well as normalized to unity, as square-integrable functions over the time domain. Also, it is straightforward to show that

$$\sum_{j=1}^{3N} \mathbf{K}^{(0)}(i,j)\mathbf{x}_j = \mathbf{x}_i^{\mathrm{rec}}. \tag{13}$$

This means that $\mathbf{K}^{(0)}$ serves a dual role both as the correlation of the LFA outputs (see above) and as the projection operator onto the low-frequency subspace spanned by n principal modes.

6.2 LFA sparsification

LFA theory replaces the n global principal modes with a much larger number $3N$ of local LFA output functions $O(i)$. Although locality was achieved, it came at a price of expanding again to the full number of degrees of freedom, $3N$. Therefore, an additional dimensionality reduction step is required in the LFA output space.

The sparsification takes advantage of the fact that neighboring outputs are highly correlated. We approximate the entire $3N$ outputs $O(i)$ with only a small subset of outputs that correspond to the strongest local features. The other $O(i)$ can then be reasonably well predicted via the correlations $\mathbf{K}^{(0)}$.

We begin with an empty set $\mathcal{M}$ of outputs. At each step, out of the N available atoms we add a seed atom, whose x, y, or z coordinate has the maximum reconstruction mean-square error, as the next member of $\mathcal{M}$, under the condition that the seed atom and its nearest neighbors are distinct from previously found atoms. This assures that the corresponding atom is decorrelated via $\mathbf{K}^{(0)}$ from the atoms corresponding to already chosen indices. We keep

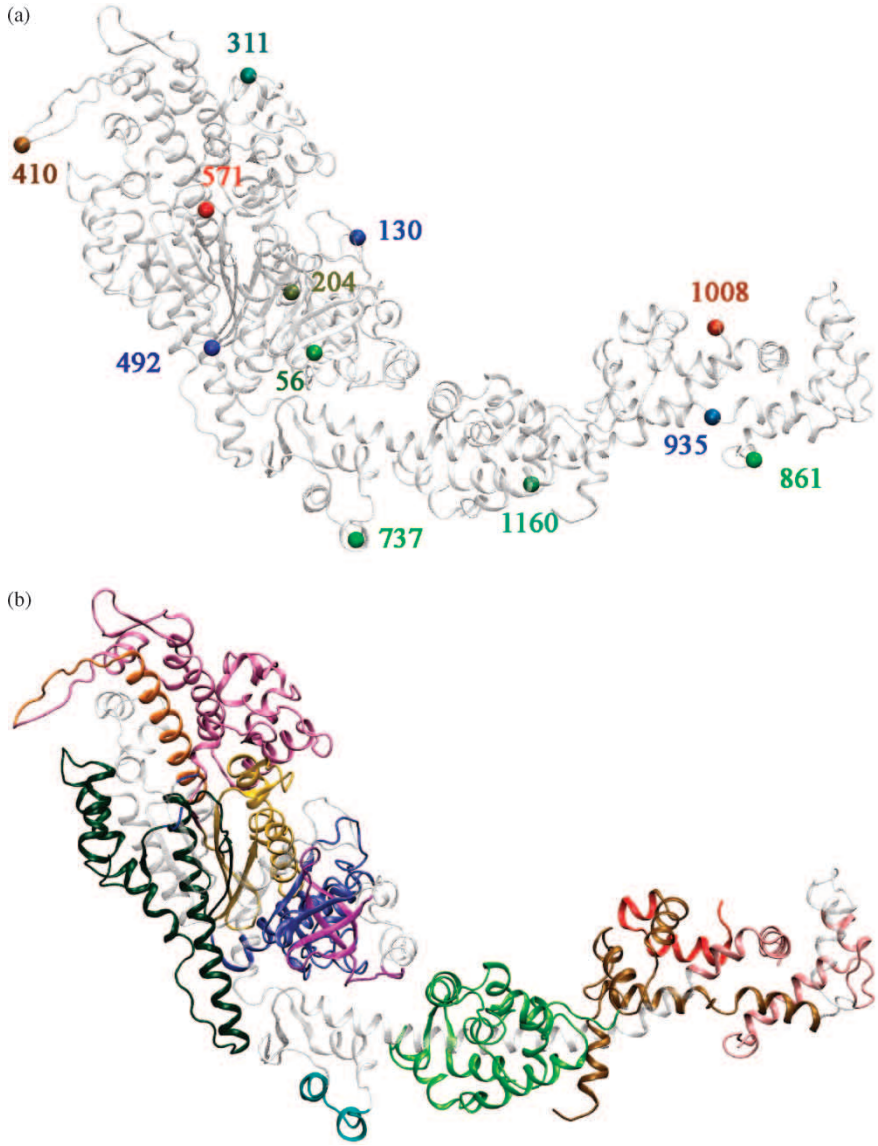

Figure 4. Application of LFA to the myosin motor domain. (a) Twelve seed atoms from the standard MD simulation labeled by residue number (table 1). The color coding indicates the order of selection in the LFA sparsification (see text). The colors vary from red to blue in an increasing order of selection. (b) The 12 corresponding local dynamic domains (random color assignment). The domains are defined as the localized, contiguous regions of positive correlation ($\mathbf{K}^{(0)} > 0$) with the corresponding seed atom.

adding seed atoms to $\mathcal{M}$ until n atoms are chosen (the entire set of $O(i)$ is reconstructed without error at this time). Complete algorithmic details of the sparsification are given elsewhere [60].

6.3 LFA results

We have applied LFA to an $n = 12$ dimensional subspace extracted with PCA from the simulation trajectories of myosin. Our goal was to construct a topographic representation of functional dynamics in terms of local features. The results are shown in figure 4. The location of the selected seed atoms at the protein surface indicate that they are allocated predominantly at the most flexible regions (figure 4(a)). The atoms are also located near functionally well know parts of the molecule. For example, Asn 410 is located in an actin binding loop, the so-called "cardiomyopathy loop", whose disruption by missense mutations is implicated in the familial hypertrophic cardiomyopathy [63]. Lys 130 is the entryway for the active site where ATP is hydrolyzed and chemical energy freed is turned into mechanical motion. Four seeds atoms (861, 935, 1008, 1160) correspond to the light chains that stabilize myosin's lever arm.

LFA represents a local feature by one seed atom and its neighboring correlated region (dynamic domain). Defining a dynamic domain as the contiguous atoms that have positive correlations with a seed atom, we have identified the prominent local features associated with the seed atoms in figure 4(b). A detailed biological interpretation of these functional "machine parts" of myosin will be given elsewhere. Suffice it to note that the mobile regions are evenly distributed across the molecule.

Table 1. LFA seed atoms obtained from MD and ACM simulations of myosin. The 12 C_α seed atoms in the MD case were reordered according to the selection order of ACM. The sequence numbering for $C_\alpha < 843$ (myosin heavy chain) corresponds to PDB entry 2MYS [66], the numbering for $C_\alpha > 859$ (myosin light chains) corresponds to the online supplementary file "motor_domain.pdb" published by Holmes *et al.* [50].

MD	*ACM*
410	409
1008	1016
861	900
571	571
737	751
56	57
1160	1140
130	129
311	295
492	497
204	204
935	667

Table 1 lists the seed atoms obtained by LFA of both MD and ACM trajectories. Five of the 12 seed atoms from both cases are at most one residue apart. It is remarkable that a significant number of features are conserved, even though the MD simulation was performed at thermal equilibrium whereas in the ACM case the conformational sampling was enhanced by selective heating of the low-frequency normal modes. Despite the overall dissipation of thermal energy from the slow to the fast modes, and the much larger conformational variability of the ACM simulation, a significant number of local features are robust enough to withstand drastic changes in the thermodynamics and conformation of the molecule.

7. Conclusions

Collective coordinate methods continue to play an important role in the dynamical analyses of nanoscale functional motion of biomolecules by providing an important dimensionality reduction. Although it is clear that individual principal or normal modes may overestimate the coherence of long-distance motions due to limited sampling or due to the required approximations of the physics, it is possible to reduce the artifacts from orthogonalization by enforcing the symmetry of a biomolecule in the analysis. The dimensionality reduction also enables a subsequent local representation of the dynamics, which provides for a significant improvement in the reproducibility and convergence of the statistical sampling and a more reliable assignment of local modes across different MD time windows [64]. In this filtering role of collective coordinates, it is not necessary to know *a priori* which particular principal or normal modes (or which linear combination of modes) are functionally relevant. The minimal assumption is that only the combined subspace is relevant, as suggested by the findings of Amadei *et al.* [33] and by a recent survey of NMA, where observed conformational changes are most often contained within the subspace of the first 12 low-frequency modes [57]. Extensions of collective coordinate methods to hybrid enhanced sampling techniques such as ACM provide interesting insights into protein folding and biomolecular dynamics on the nanometer scale, and suggests new approaches to the refinement and interpretation of experimental results in structural biology.

Acknowledgements

M.S. and D.C.S. gratefully acknowledge Professor Lydia Kavraki and Dr Mark Moll from Department of Computer Science at Rice U. for introduction to the symmetry problem and for providing computations on HIV-1 protease. This work is supported in part by grants from NIH (1R01GM62968), Human Frontier Science Program (RGP0026/2003), Alfred P. Sloan Foundation (BR-4297) to W.W., a training fellowship from the Keck Center Pharmacoinformatics Training Program of the Gulf Coast Consortia (NIH Grant 1R90DK071505) to Z.Z., as well as NSF Grants CCR-0306503 and ACI-0325081 to D.C.S.

References

[1] E. Jakobsson. The top ten advances of the last decade and the top ten challenges of the next. *Biomed. Comput. Rev.*, **1**(1), 11 (2005).
[2] B. Alberts. The cell as a collection of protein machines: preparing the next generation of molecular biologists. *Cell*, **92**, 291 (1998).
[3] M. Karplus. Molecular dynamics: applications to proteins. In *Modelling of Molecular Structures and Properties*, Volume 71 of *Studies in Physical and Theoretical Chemistry.* Proceedings of an International Meeting, J.-L. Rivail (Ed.), pp. 427–461, Elsevier Science Publishers, Amsterdam (1990).
[4] C.L. Brooks III, M. Karplus, B.M. Pettitt. *Proteins: A Theoretical Perspective of Dynamics, Structure and Thermodynamics,* Volume LXXI of *Advances in Chemical Physics*, John Wiley & Sons, New York (1988).
[5] J.B. Clarage, T. Romo, B.K. Andrews, B.M. Pettitt, G.N. Phillips. A sampling problem in molecular dynamics simulations of macromolecules. *Proc. Natl. Acad. Sci. USA*, **92**, 3288 (1995).
[6] G. Hernández, F.E. Jenney Jr., M.W.W. Adams, D.M. LeMaster. Millisecond time scale conformational flexibility in a hyperthermophile protein at ambient temperature. *Proc. Natl. Acad. Sci. USA*, **97**, 3166 (2000).
[7] J.J. Falke. A moving story. *Science*, **295**, 1480 (2002).
[8] P. Bjorkman, J. Abelson, J. Kobori. *Supramolecular Assemblies: Current Technology and Resource Needs*, The Agouron Institute, Pasadena, (1999).
[9] J.-P. Ryckaert, G. Ciccotti, H.J.C. Berendsen. Numerical integration of the Cartesian equations of motion of a system with constraints: molecular dynamics of *n*-alkanes. *J. Comp. Phys.*, **23**, 327 (1977).
[10] C.S. Peskin, T. Schlick. Molecular dynamics by the backward Euler method. *Commun. Pure Appl. Math.*, **42**, 1001 (1989).
[11] G. Zhang, T. Schlick. The Langevin/implicit-Euler/normal-mode scheme for molecular dynamics at large time steps. *J. Chem. Phys.*, **101**, 4995 (1994).
[12] M. Watanabe, M. Karplus. Simulations of macromolecules by multiple-time-step methods. *J. Phys. Chem.*, **99**(15), 5680 (1995).
[13] P. Eastman, S. Doniach. Multiple time step diffusive Langevin dynamics for proteins. *Proteins Struct. Funct. Genet.*, **30**, 215 (1998).
[14] K.A. Feenstra, B. Hess, H.J.C. Berendsen. Improving efficiency and accuracy of molecular dynamics simulations of hydrogen-rich systems. *J. Comput. Chem.*, **20**, 786 (1999).
[15] A.K. Mazur, V.E. Dorofeev, R.A. Abagyan. Derivation and testing of explicit equations of motion for polymers described by internal coordinates. *J. Comp. Phys.*, **92**, 261 (1991).

[16] C.D. Schwieters, G.M. Glore. Internal coordinates for molecular dynamics and minimization in structure determination and refinement. *J. Magn. Reson.*, **152**, 288 (2001).

[17] Y. Zhou, M. Karplus. Interpreting the folding kinetics of helical protein. *Nature*, **401**, 400 (1999).

[18] C. Clementi, H. Nymeyer, J.N. Onuchic. Topological and energetic factors: what determines the structural details of the transition state ensemble and "en-route" intermediates for protein folding? An investigation for small globular proteins. *J. Mol. Biol.*, **298**, 937 (2000).

[19] I. Bahar, B. Erman, T. Haliloglu, R.L. Jernigan. Efficient characterization of collective motions and interresidue correlations in proteins by low-resolution simulations. *Biochemistry*, **36**, 13512 (1997).

[20] A. Sali, E.I. Shakhnovich, M. Karplus. Kinetics of protein folding: a lattice model study of the requirement for folding to the native state. *J. Mol. Biol.*, **235**, 1614 (1994).

[21] H. Li, N. Winfreen, C. Tang. Emergency of preferred structures in a simple model of protein folding. *Science*, **273**, 666 (1996).

[22] E.I. Shakhnovich. Theoretical studies of protein-folding thermodynamics and kinetics. *Curr. Opin. Struct. Biol.*, **7**, 29 (1997).

[23] H. Grubmüller. Predicting slow structural transitions in macromolecular systems: conformational flooding. *Phys. Rev. E*, **52**(3), 2893 (1995).

[24] C. Bartels, M. Karplus. Probability distributions for complex systems: adaptive umbrella sampling of the potential energy. *J. Phys. Chem. B*, **102**, 865 (1998).

[25] N. Nakajima, H. Nakamura, A. Kidera. Multicanonical ensemble generated by molecular dynamics simulations for enhanced conformational sampling of peptides. *J. Phys. Chem. B*, **101**, 817 (1997).

[26] A.M. Ferrenberg, R.H. Swendsen. New Monte Carlo technique for studying phase transitions. *Phys. Rev. Lett.*, **61**, 2635 (1988).

[27] X. Wu, S. Wang. Self-guided molecular dynamics simulation for efficient conformational search. *J. Phys. Chem. B*, **102**, 7238 (1998).

[28] J. Schlitter, M. Engels, P. Krüger, E. Jacoby, A. Wollmer. Targeted molecular dynamics simulation of conformational change—application to the T ↔ R transition in insulin. *Mol. Simulation*, **10**(2–6), 291 (1993).

[29] W. Wriggers, K. Schulten. Investigating a back door mechanism of actin phosphate release by steered molecular dynamics. *Proteins Struct. Funct. Genet.*, **35**, 262 (1999).

[30] Z. Zhang, Y. Shi, H. Liu. Molecular dynamics simulations of peptides and proteins with amplified collective motions. *Biophys. J.*, **84**, 3583 (2003).

[31] T. Horiuchi, N. Go. Projection of Monte Carlo and molecular dynamics trajectories onto the normal mode axes: human lysozyme. *Proteins Struct. Funct. Genet.*, **10**, 106 (1991).

[32] A. Kitao, F. Hirata, N. Go. The effects of solvent on the conformation and the collective motions of proteins: normal mode analysis and molecular dynamics simulations of Melittin in water and in vacuum. *Chem. Phys.*, **158**, 447 (1991).

[33] A. Amadei, A.B.M. Linnsen, H.J.C. Berendsen. Essential dynamics of proteins. *Proteins Struct. Funct. Genet.*, **17**, 412 (1993).

[34] M.F. van Aalten, A. Amadei, A.B.M. Linssen, V.G.H. Eijsink, G. Vriend, H.J.C. Berendsen. The essential dynamics of thermolysin: conformation of the hinge-bending motion and comparison of simulations in vacuum and water. *Proteins Struct. Funct. Genet.*, **22**, 45 (1995).

[35] A.E. García. Large-amplitude nonlinear motions in proteins. *Phys. Rev. Lett.*, **68**, 2696 (1992).

[36] A. Kitao, N. Go. Investigating protein dynamics in collective coordinate space. *Curr. Opin. Struct. Biol.*, **9**, 164 (1999).

[37] H.J.C. Berendsen, S. Hayward. Collective protein dynamics in relation to function. *Curr. Opin. Struct. Biol.*, **10**, 165 (2000).

[38] B.R. Brooks, D. Janezic, M. Karplus. Harmonic analysis of large systems I. Methodology. *J. Comput. Chem.*, **16**, 1522 (1995).

[39] D.A. Case. Normal mode analysis of protein dynamics. *Curr. Opin. Struct. Biol.*, **4**, 285 (1994).

[40] K. Karhunen. Über lineare Methoden in der Wahrscheinlichkeitsrechnung. *Ann. Acad. Sci. Fenn A*, 137 (1947).

[41] D. Janezic, R.M. Venable, M. Karplus. Harmonic analysis of large systems III. Comparison with molecular dynamics. *J. Comput. Chem.*, **16**, 1554 (1995).

[42] M. Karplus, J.N. Jushick. Method for estimating the configurational entropy of macromolecules. *Macromolecules*, **14**, 325 (1981).

[43] R.M. Levy, A.R. Srinivasan, W.K. Olson, J.A. McCammon. Quasiharmonic method for studying very low frequency modes in proteins. *Biopolymers*, **23**, 1099 (1984).

[44] D.A. Case. Normal mode analysis of biomolecular dynamics. In *Computer Simulation of Biomolecular Systems*, W.F. van Gunsteren, P.K. Weiner, A.J. Wilkinson (Eds), 3, pp. 284–301, Kluwer Academic Publishers, Dordrecht, Netherlands (1997).

[45] D.A. McQuarrie. *Statistical Mechanics*, Harper and Row, New York, CA (1976).

[46] A. Kitao, S. Hayward, N. Go. Energy landscape of a native protein: jumping-among-minima model. *Proteins Struct. Func. Genet.*, **33**, 496 (1998).

[47] M.A. Balsera, W. Wriggers, Y. Oono, K. Schulten. Principal component analysis and long time protein dynamics. *J. Phys. Chem.*, **100**(7), 2567 (1996).

[48] I. Bahar, A.R. Atilgan, B. Erman. Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential. *Fold. Des.*, **2**, 173 (1997).

[49] P. Chacón, F. Tama, W. Wriggers. Mega-Dalton biomolecular motion captured from electron microscopy reconstructions. *J. Mol. Biol.*, **326**, 485 (2003).

[50] K.C. Holmes, I. Angert, F.J. Kull, W. Jahn, R.R. Schröder. Electron cryo-microscopy shows how strong binding of myosin to actin releases nucleotide. *Nature*, **425**, 423 (2003).

[51] W.R.P. Scott, P.H. Hünenberger, I.G. Tironi, A.E. Mark, S.R. Billeter, J. Fennen, A.E. Torda, T. Huber, P. Krüger, W.F. van Gunsteren. The GROMOS biomolecular simulation program package. *J. Phys. Chem. A*, **103**, 3596 (1999).

[52] H.J.C. Berendsen, J.P.M. Postma, W.F. van Gunstreren, J. Herrnans. Interaction models for water in relation to protein hydration. In *Intermolecular Forces*, B. Pullman (Ed.), pp. 331–342, Reidel, Dordrecht, Netherlands (1981).

[53] H.J.C. Berendsen, J.P.M. Postma, W.F. van Gunsteren, A. DiNola, J.R. Haak. Molecular dynamics with coupling to an external bath. *J. Chem. Phys.*, **81**(8), 3684 (1984).

[54] S.M. Block. Fifty ways to love your lever: myosin motors. *Cell*, **87**, 151 (1996).

[55] J.D. Jontes, E.M. Wilson-Kubalek, R.A. Milligan. A 32° tail swing in brush border myosin I on ADP release. *Nature*, **378**, 751 (1995).

[56] K. Kitamura, M. Tokunaga, A.H. Iwane, T. Yanagida. A single myosin head moves along an actin filament with regular steps of 5.3 nanometres. *Nature*, **397**, 129 (1999).

[57] F. Tama, Y.-H. Sanejouand. Conformational change of proteins arising from normal mode calculations. *Protein Eng.*, **14**, 1 (2001).

[58] H. Brandstetter, J.S. Kim, M. Groll, R. Huber. Crystal structure of the tricorn protease reveals a protein disassembly line. *Nature*, **414**, 466 (2001).

[59] M. Shah, D.C. Sorensen. A symmetry preserving singular value decomposition. Technical Report TR05-01, Dept. Computational and Applied Math., Rice University, Houston, 2005. Available at http://www.caam.rice.edu.

[60] Z. Zhang, W. Wriggers. Local feature analysis: a statistical theory for reproducible essential dynamics of large macromolecules. *Proteins Struct. Funct. Bioinformatics*, (2006) In Press.

[61] R.B. Lehoucq, D.C. Sorensen, C. Yang. *ARPACK Users' Guide — Solution of Large-Scale Eigenvalue Problems with Implicitly Restarted Arnoldi Methods*, SIAM, Philadelphia (1998).

[62] J. Erickson, D.J. Neidhart, J. VanDrie, D.J. Kempf, X.C. Wang, D.W. Norbeck, J.J. Plattner, J.W. Rittenhouse, M. Turon, N. Wideburg, et al. Design, activity, and 2.8 A crystal structure of a C2 symmetric inhibitor complexed to HIV-1 protease. *Science*, **349**, 527 (1990).

[63] L. Fananapazir, M.C. Dalakas, F. Cyran, G. Cohn, N.D. Epstein. Missense mutations in the β-myosin heavy-chain gene cause central core disease in hypertrophic cardiomyopathy. *Proc. Natl. Acad. Sci. USA*, **90**, 3993 (1993).

[64] L. Zhang, J. Hermans. Hydrophilicity of cavities in proteins. *Proteins Struct. Funct. Genet.*, **24**, 433 (1996).

[65] W.F. Humphrey, A. Dalke, K. Schulten. VMD — visual molecular dynamics. *J. Mol. Graph.*, **14**, 33 (1996).

[66] I. Rayment, W.R. Rypniewski, K. Schmidt-Bäse, R. Smith, D.R. Tomchick, M.M. Benning, D.A. Winkelmann, G. Wesenberg, H. M. Holden. Three-dimensional structure of myosin subfragment-1: a molecular motor. *Science*, **261**, 50 (1993).