J Multidisciplinary Scientific Journal

MDPI

Review

Principal Component Analysis and Related Methods for Investigating the Dynamics of Biological Macromolecules

Akio Kitao

School of Life Science and Technology, Tokyo Institute of Technology, Tokyo 152-8550, Japan; akitao@bio.titech.ac.jp

Abstract: Principal component analysis (PCA) is used to reduce the dimensionalities of high-dimensional datasets in a variety of research areas. For example, biological macromolecules, such as proteins, exhibit many degrees of freedom, allowing them to adopt intricate structures and exhibit complex functions by undergoing large conformational changes. Therefore, molecular simulations of and experiments on proteins generate a large number of structure variations in high-dimensional space. PCA and many PCA-related methods have been developed to extract key features from such structural data, and these approaches have been widely applied for over 30 years to elucidate macromolecular dynamics. This review mainly focuses on the methodological aspects of PCA and related methods and their applications for investigating protein dynamics.

Keywords: principal component analysis; collective variables; molecular dynamics; energy landscape; solvent effects; linear response theory; independent component analysis

check for updates

Citation: Kitao, A. Principal Component Analysis and Related Methods for Investigating the Dynamics of Biological Macromolecules. *J* **2022**, *5*, 298–317. https://doi.org/10.3390/j5020021

Academic Editor: Johan Jacquemin

Received: 11 May 2022
Accepted: 17 June 2022
Published: 20 June 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

1. Historical Overview

Principal component analysis (PCA) is a widely used multivariate analysis approach, originally proposed about 100 years ago [1,2], that has found increasing applications since the widespread availability of digital computers to reduce the dimensionality of high-dimensional datasets. This reduction is enabled by linear transformation from the original variables to new collective variables, so that a small number of "principal components" dominate the features of the dataset. Now PCA is considered as an unsupervised machine learning technique.

The structures of proteins and other biological macromolecules are well characterized by a set of multidimensional variables, such as atomic coordinates and dihedral angles, and information regarding the dynamics of these molecules is typically obtained as a time series of high-dimensional data or an ensemble of experimentally determined structures. Although such large ensembles of high-dimensional data contain useful information, they are not easily interpretable. Therefore, extracting important features from high-dimensional data is essential to understand the dynamics of biological macromolecules.

Similar to the increased application of PCA in other areas, the use of PCA in analyzing protein dynamics has gradually become more common as the performance of computers has improved, making the molecular simulations of proteins more accessible. The first molecular dynamics (MD) simulation of a small folded protein, bovine pancreatic trypsin inhibitor (BPTI), in vacuum was conducted in 1977 [3], and the first protein normal mode analysis (NMA) for BPTI was performed in 1983 [4–6]. NMA is a harmonic approximation of protein dynamics at a potential energy minimum, and clearly shows that the low-frequency normal modes of proteins are collective motions of the atoms spread over the entire protein (namely, global motions) and that the lowest normal mode frequency is a few cm^{-1}. Since the vibrational frequencies of bond stretching modes are higher than this by three orders of magnitude, the amplitudes of the lowest and highest modes also differ by three-fold, indicating the highly anisotropic nature of proteins even within the

range of vibrational motions. This high anisotropy may be partly attributed to the highly packed structures of folded native proteins, whose packing densities are comparable to that of a face-centered cubic lattice [7]. In highly packed structures, local motions uncorrelated with the surroundings are limited to small amplitudes because of possible collisions, while concerted motions of groups of atoms such as protein domains or loops can move in certain directions largely without altering atomic packing. In 1981, Karplus and Kushick proposed a method to estimate the configurational entropy of macromolecules from NMA, MD and Monte Carlo (MC) simulations [8]. That publication also showed that simulations with (NMA) and without harmonic approximation (MD and MC) can be connected by PCA. The length of the first reported protein MD was 8.8 ps [3], which roughly corresponds to one period of the lowest-frequency normal mode of typical small globular proteins and thus was insufficiently long to sample large-amplitude motions of the protein. However, increasing simulation lengths allowed investigation of the quasi-harmonic features of butane and BPTI, mainly focusing on quasi-harmonic frequencies deduced from PCA [9,10]. Later, projecting simulation trajectories onto collective coordinates was shown to be very useful for characterizing dominant protein dynamics, but the early stages of this endeavor used low-frequency normal modes for the projected collective variables [11]. Since normal modes are determined based only on one energy minimum, they are not necessarily the best choice to investigate the anharmonic nature of protein dynamics. In contrast, PCA determines principal coordinates as the collective coordinates, which incorporate anharmonic features included in the MD or MC trajectory. Longer and more realistic MD simulations in solution were performed from the 1980s to the early 1990s, allowing the PCA of MD trajectories. In the early 1990s, the anisotropic and anharmonic nature of native protein dynamics was elucidated by PCA, focusing on principal components (PCs), defined as the projections onto the principal coordinates [12–15]. PCA was also shown to be useful for analyzing simulation trajectories of protein folding/non-folding dynamics [16,17]. The past three decades have been seen the frequent use of PCA to investigate the dynamic behavior of biopolymers, as well as many important methodological improvements and the elucidation of simulated dynamic features [18–23]. Since PCA employs a variance–covariance matrix for dimensionality reduction, it is useful to characterize large-amplitude conformational change in molecules, such as protein domain motion and folding. However, PCA may not be sensitive for detecting localized, small amplitude but functionally important motions, such as backrub motion [24], peptide-plane flip [25], the side-chain flip and path-preserving motions [26].

This review provides an overview of PCA and related methods and their applications for investigating protein dynamics, focusing mainly on methodological aspects. In addition, some basic concepts and important findings obtained during the early years of this field are revisited for the benefit of non-experts, as well as a review of the latest progress in PCA-related research. The following PCA applications demonstrate the examples in which macromolecular dynamics cannot be well characterized without the use of PCA.

2. Basic Concept behind PCA

The investigation of macromolecular dynamics by PCA requires the selection of certain degrees of freedom of the target molecule that characterize the dynamics well. Consider a vector of general coordinates of a target molecule or molecules, $\mathbf{q}$, and suppose that $\mathbf{q} = \{q_i\}$ is a column vector consisting of f variables ($i = 1, \ldots, f$). Thus, $\langle \mathbf{q} \rangle = \{\langle q_i \rangle\}$ is the average and $\Delta\mathbf{q} = \mathbf{q} - \langle \mathbf{q} \rangle$ is the deviation from the average. To explicitly indicate an index of the mth data point among M ($m = 1, \ldots, M$), we use the expression $\Delta\mathbf{q}_m$. When the MD trajectory is considered, the expression $\Delta\mathbf{q}(t_m)$ is employed to specify a coordinate set at time t_m. To conduct PCA, a variance–covariance matrix $\mathbf{C}$ is introduced:

$$\mathbf{C} = \frac{1}{M}\sum_{m=1}^{M} \Delta\mathbf{q}_m \, \Delta\mathbf{q}_m^T = \left\langle \Delta\mathbf{q} \, \Delta\mathbf{q}^T \right\rangle = \left\{ \left\langle (q_i - \langle q_i \rangle)(q_j - \langle q_j \rangle) \right\rangle \right\}, \tag{1}$$

where $\Delta\mathbf{q}^T$ represents the transpose of $\Delta\mathbf{q}$ and $\langle \cdots \rangle$ denotes the simple average over a given dataset. The matrix $\mathbf{C} = \left\{C_{ij}\right\}$ is a positive semidefinite whose eigenvalues are non-negative. By introducing an $f \times M$ matrix of the whole dataset:

$$\mathbf{Q} = \{\Delta\mathbf{q}_1 \cdots \Delta\mathbf{q}_M\}, \tag{2}$$

A can be obtained as a matrix product:

$$\mathbf{C} = \frac{1}{M}\mathbf{Q}\mathbf{Q}^T, \tag{3}$$

where $\mathbf{Q}^T$ denotes the transpose of $\mathbf{Q}$. By solving the standard eigenvalue problem with the orthonormal condition:

$$\mathbf{C}\mathbf{V} = \mathbf{V}\boldsymbol{\lambda}, \tag{4}$$

$$\mathbf{V}\mathbf{V}^T = \mathbf{V}^T\mathbf{V} = \mathbf{I}, \tag{5}$$

we obtain $\mathbf{V}$, $\boldsymbol{\lambda}$ and $\mathbf{I}$, which are the eigenvector, eigenvalue and unit matrices, respectively. $\boldsymbol{\lambda}$ is a diagonal matrix whose αth diagonal element $\{\lambda_\alpha\}$ ($\alpha = 1, \ldots, f$) is the variance of the αth PC, and the αth column vector $\mathbf{v}_\alpha$ of $\mathbf{V}$ is the corresponding eigenvector. Typically, λ_α is sorted in descending order such that the first PC shows the largest variance λ_1 and the corresponding column vector $\mathbf{v}_1$ of $\mathbf{V} = \left(\mathbf{v}_1 \cdots \mathbf{v}_f\right)$ indicates the eigenvector of the first PC.

Projection of $\Delta\mathbf{q}_m$ onto $\mathbf{V}$ provides:

$$\boldsymbol{\sigma}_m = \mathbf{V}^T\Delta\mathbf{q}_m, \tag{6}$$

where $\boldsymbol{\sigma}_m$ is a column vector of the projections (principal components). The overall linear transformation of $\mathbf{Q}$ using $\mathbf{V}$ gives a projection matrix $\boldsymbol{\Sigma} = \left(\boldsymbol{\sigma}_1 \cdots \boldsymbol{\sigma}_f\right)$ onto the PC:

$$\boldsymbol{\Sigma} = \mathbf{V}^T\mathbf{Q}. \tag{7}$$

The $f \times M$ matrix $\boldsymbol{\Sigma}$ represents the matrix of principal components, which are collective variables defined as a linear combination of the original coordinates and the elements of $\mathbf{V}$ are coefficients for the transformation.

PCA can be conducted by solving the standard eigenvalue problem (Equation (4)), which requires the diagonalization of $\mathbf{C}$ ($f \times f$ matrix). PCA can also be performed by the singular value decomposition (SVD) of $\mathbf{Q}$ ($f \times M$ matrix). SVD directly provides a decomposition of $\mathbf{Q}$ into three matrices:

$$\mathbf{Q} = \sqrt{M}\mathbf{V}\boldsymbol{\lambda}^{1/2}\,\mathbf{U}^T, \tag{8}$$

where $\mathbf{U}^T$ is an $M \times M$ matrix of normalized projections defined as:

$$\mathbf{U}^T = \frac{1}{\sqrt{M}}\boldsymbol{\lambda}^{-1/2}\boldsymbol{\Sigma} = \frac{1}{\sqrt{M}}\boldsymbol{\lambda}^{-1/2}\,\mathbf{V}^T\mathbf{Q}. \tag{9}$$

In this case, $\boldsymbol{\lambda}^{1/2}$ is an $f \times M$ matrix whose non-diagonal elements are zero. From Equations (4), (5) and (9), it is straightforward to obtain the condition:

$$\mathbf{U}^T\mathbf{U} = \mathbf{I}. \tag{10}$$

To quantify the usefulness of PCA for a target dataset, the PC contribution to the total variance is examined, which is defined as:

$$\chi_\alpha = \frac{\lambda_\alpha}{\sum_{\alpha=1}^{f} \lambda_\alpha}. \tag{11}$$

If the contributions from a small number of PCs to the total variance are dominant, PCA is useful for dimensionality reduction because most of the total variance originates from these PCs. Typically, folded proteins are highly anisotropic in nature and PCA is thus useful for analyzing protein dynamics [18–20,22].

It is also straightforward to consider hierarchy in the distribution of a target dataset, for example, the distribution of clusters and the distributions of data in each cluster. Suppose that the dataset of interest is clustered into L groups, each consisting of n_l data points. In this case, $\langle \cdots \rangle_l$ indicates the mean over n_l and $f_l = n_l/K$ is the fraction of the lth group. The variance–covariance matrix can be divided into two terms as [27]:

$$\begin{aligned} \mathbf{C} &= \mathbf{C}^{\mathrm{JAM}} + \mathbf{C}^{\mathrm{intra}} \\ &= \left\{ \sum_{l=1}^{L} f_l \left\langle (\langle q_i \rangle_l - \langle q_i \rangle) \left(\langle q_j \rangle_l - \langle q_j \rangle \right) \right\rangle + \sum_{l=1}^{L} f_l \left\langle (q_i - \langle q_i \rangle_l) \left(q_j - \langle q_j \rangle_l \right) \right\rangle_l \right\} \end{aligned} \tag{12}$$

The first term represents the variance–covariance originating from the distribution of the means of the groups and the second term shows the f_l weighted average of intra-group variance–covariance. Although Equation (12) shows a two-hierarchy model, extension to multiple hierarchy is straightforward. If the distribution of each group shows fluctuations in a certain local energy minimum, the first term originates from jumping among energy minima. Using this formulation, the jumping-among-minima (JAM) model shows that the JAM motions that contribute to the first term dominate a small number of anharmonic large-amplitude motions and the second term is attributed to nearly harmonic fluctuations around local energy minima that are detected as Gaussian-like distributions [19,27]. This hierarchical view of protein collective dynamics is useful for understanding the boson peak and glass transition of proteins [28], and for identifying multiple conformers in nuclear magnetic resonance (NMR)-derived structure ensembles of proteins [29,30].

3. Error in PCA

A dataset to be analyzed by PCA may contain an insufficient number of samples statistically, which can result in instability of the obtained eigenvectors. This situation frequently occurs when standard MD simulations of macromolecules are conducted with all-atom models because the accessible simulation time scale tends to be shorter than the characteristic time scale of macromolecular movement. Hess showed that random diffusions in high-dimensional space can result in cosine-shaped projections of the first few dominant PCs, indicating that short simulation trajectories should be carefully treated [31]. However, using random matrix theory (RMT) [32,33], Palese showed that protein dynamics is not truly Brownian even on a short time scale (~1 ns) while PCA leads to cosine-shaped projections, using apo Cox17 as a model protein [34]. In this method, $\mathbf{C}$ is considered to be the sum of the random component $\mathbf{C}_r$ and non-random component $\mathbf{C}_{nr}$. $\mathbf{C}_r$ is determined by an iterative method based on RMT, providing the cleaned $\mathbf{C}_{nr}$ without the random component and its eigenvectors. Palese also proposed random component analysis (RCA) as a random projection (RP) algorithm [35]. In RCA, PCA of the correlation matrix $\mathbf{P} = \left\{ C_{ij} / \sqrt{C_{ii} C_{jj}} \right\}$ is considered, and $\mathbf{P}$ is replaced by a random symmetric correlation matrix $\mathbf{M}$ as a dummy correlation. PCA of $\mathbf{M}$ and projection onto the obtained eigenvectors

provides the random components. RCA provides dimensionality reduction and cluster detection comparable to that of PCA. Ref. [36] introduced and examined a parameter that evaluates the overlap between such subspaces, called the root mean squared inner product (RMSIP), and suggested that PCA for the concatenated equivalent trajectories achieves better reproducibility.

4. Relation with NMA

NMA is closely related to PCA as mentioned in Section 1. To conduct NMA, the second derivative matrix of potential energy E (Hessian) $\mathbf{F}$ should be calculated at a certain conformation, typically at a local potential energy minimum conformation where the first derivatives $\partial E/\partial q_i = 0$ for all i:

$$\mathbf{F} = \{f_{ij}\} = \left\{\partial^2 E/\partial q_i \partial q_j\right\} \tag{13}$$

when Cartesian coordinates are used for Equation (13), $\mathbf{q}$ should be mass weighted Cartesian coordinates. To obtain normal mode frequencies and eigenvectors, the standard eigenvalue problem of $\mathbf{F}$ is solved as:

$$\mathbf{FW} = \mathbf{W}\boldsymbol{\omega}^2, \tag{14}$$

$$\mathbf{W}^T\mathbf{W} = \mathbf{W}\mathbf{W}^T = \mathbf{I}. \tag{15}$$

The βth column vector of $\mathbf{W} = \left(\mathbf{w}_1 \cdots \mathbf{w}_f\right)$, $\mathbf{w}_\beta$, represents the βth eigenvector. The eigenvalue matrix $\boldsymbol{\omega}^2 = \left\{\omega_\beta^2\right\}$ determines the angular frequency of the normal modes ω_β. Since the variance–covariance matrix $\mathbf{C}$ is related to $\mathbf{F}$ by $\mathbf{C} = k_B T\mathbf{F}^{-1}$ (k_B: Boltzmann constant, T: absolute temperature), we obtain the relation for the harmonic system:

$$\lambda_\beta = k_B T/\omega_\beta^2 \tag{16}$$

Therefore, comparing λ_α obtained by PCA of MD or MC trajectories to the NMA-derived $k_B T/\omega_\alpha^2$ is a straightforward way to examine the anharmonicity or quasi-harmonic features of protein dynamics. $\mathbf{W}$ is determined for a potential energy minimum, while MD simulation can sample multiple energy minima. Therefore, $\mathbf{V}$ obtained from an MD trajectory can be significantly different from $\mathbf{W}$ calculated around a particular local energy minimum. This difference becomes larger as the MD length is increased. To consider the difference between two collective variables and to examine the anharmonicity of an energy surface, the variance expected from NMA along the αth PC is obtained by:

$$\lambda_\alpha^{har} = k_B T\sum_\beta \frac{\left(\mathbf{w}_\beta \cdot \mathbf{v}_\alpha\right)^2}{\omega_\beta^2}. \tag{17}$$

λ_α^{har} is further used to define the anharmonicity observed in MD along the αth PC, namely, the anharmonicity factor:

$$\mu_\alpha = \left(\lambda_\alpha/\lambda_\alpha^{har}\right)^{1/2}. \tag{18}$$

μ_α is unity if the variance is equal to that expected from NMA, indicating that the energy surface along the αth PC is nearly harmonic [27,37]. For short MD trajectories up to 1 ns, the majority of PCs are harmonic and less than 1% of PCs show $\mu_\alpha > 2$, and anharmonic motions dominantly contribute to the total variance [19,27].

NMA of proteins was originally conducted with full atomic models [4–6]. Over the past 20 years or so, NMA has been more widely used with coarse-grained models as well as with atomic models. Although this is an interesting topic related to PCA, NMA is beyond the scope of this review but is covered in reviews on NMA [38–43].

5. Solvent and Other Environmental Effects on Macromolecular Dynamics

To consider solvent effects on macromolecular dynamics, one of the simplest models is the consideration of the independent Langevin equation for each PC $\sigma_\alpha(t)$ with harmonic potential:

$$\ddot{\sigma}_\alpha(t) = -\omega_\alpha^2 \sigma_\alpha(t) - \gamma_\alpha \dot{\sigma}_\alpha(t) + R_\alpha(t), \tag{19}$$

where the first term of the right-hand side shows the harmonic force and γ_α and $R_\alpha(t)$ indicate the Stokes friction coefficient and random force acting on the αth PC, respectively.

The autocorrelation functions of $\sigma_\alpha(t)$ and velocity $\dot{\sigma}_\alpha(t)$ are given by:

$$\langle \sigma_\alpha(t)\sigma_\alpha(0)\rangle = \frac{k_B T}{\omega_\alpha^2}\left[\frac{\gamma_\alpha + \varpi_\alpha}{2\varpi_\alpha}\exp\left(-\frac{\gamma_\alpha - \varpi_\alpha}{2}t\right) - \frac{\gamma_\alpha - \varpi_\alpha}{2\varpi_\alpha}\exp\left(-\frac{\gamma_\alpha + \varpi_\alpha}{2}t\right)\right], \tag{20}$$

$$\langle \dot{\sigma}_\alpha(t)\dot{\sigma}_\alpha(0)\rangle = k_B T\left[\frac{-\gamma_\alpha + \varpi_\alpha}{2\varpi_\alpha}\exp\left(-\frac{\gamma_\alpha - \varpi_\alpha}{2}t\right) + \frac{\gamma_\alpha + \varpi_\alpha}{2\varpi_\alpha}\exp\left(-\frac{\gamma_\alpha + \varpi_\alpha}{2}t\right)\right], \tag{21}$$

where $\varpi_\alpha = \sqrt{\gamma_\alpha^2 - 4\omega_\alpha^2}$. If $\gamma_\alpha > 2\omega_\alpha$, ϖ_α is a real number and the correlation functions simply consist of two exponential decays, resulting in overdamping motion. If $\gamma_\alpha < 2\omega_\alpha$, ϖ_α consists of real and imaginary parts, with the former showing a single exponential decay and the latter providing the sum of cosine and sine terms, causing damped oscillation. Because of this relation, large-amplitude motions, i.e., low-frequency motions, tend to be $\gamma_\alpha > 2\omega_\alpha$ and overdamped [12,15]. The density of state (DOS) for this system is obtained as the spectrum of the velocity correlation (Equation (21)):

$$S(\omega) = \frac{k_B T}{\pi}\frac{\gamma_\alpha \omega^2}{\left(\omega_\alpha^2 - \omega^2\right)^2 + \gamma_\alpha^2 \omega^2}. \tag{22}$$

As the ratio $\gamma_\alpha/\omega_\alpha$ becomes larger, the peak of the spectrum decreases and the area of the spectrum is shifted to the high-frequency region, resulting in lowered spectral density around the peak ω_α and seeming disappearance of the peak in DOS [12]. This model also explains why the DOS of BPTI from neutron scattering is lower than that expected from the frequency distribution in the low-frequency region < ~20 cm^{-1} [15].

Assuming each PC is a harmonic Langevin oscillator, the values of ω_α and γ_α can be estimated from the MD-derived time correlation function or its spectrum in different ways, but the obtained results can be method dependent. Considering Equation (21), one way to calculate γ_α is the time derivative of the velocity autocorrelation function at $t = 0$ [12,15]:

$$\gamma_\alpha = -\frac{1}{k_B T}\left(\frac{d}{dt}\langle \dot{\sigma}_\alpha(\Delta t)\dot{\sigma}_\alpha(0)\rangle\right)_{\Delta t = 0}. \tag{23}$$

If γ_α is calculated as the numerical derivative from a difference in the velocity correlation functions between $t = 0$ and Δt, the numerical error should be carefully considered [15]. From Equation (21), the first-order approximation of the derivative for small t is obtained as:

$$\gamma_\alpha(\Delta t) = -\frac{1}{k_B T}\frac{d}{dt}\langle \dot{\sigma}_\alpha(\Delta t)\dot{\sigma}_\alpha(0)\rangle = \gamma_\alpha + \left(\omega_\alpha^2 - \gamma_\alpha^2\right)\Delta t, \tag{24}$$

which indicates a parabolic behavior of "apparent" friction coefficients as a function of ω_α deduced by this method, and such behavior is indeed observed [12,15]. From the parabolic feature of $\gamma_\alpha(\Delta t)$, the "true" friction coefficient $\gamma_\alpha = \gamma_\alpha(0)$ after correction based on Equation (24) can be considered to be almost constant for large-amplitude PCs for a given protein [15]. The value of γ_α is quite dependent on protein size, consistent with Stokes–Einstein law, $\gamma = 6\pi a\eta/M \propto M^{-2/3}$ (a: radius, η: viscosity, M: mass) [27].

The MD-derived friction coefficient does not necessarily originate from solvent and it can be estimated for MD conducted in vacuum. Ref. [44] used fitting Equation (21) with MD-derived data for friction calculations and showed the frequency dependence of the friction

coefficient. Ref. [44] also reported that friction in vacuum is directly proportional to the intra-protein interaction of the collective mode but is also proportional to the accessible surface area of the mode in solution. In Ref. [45], MD simulations of myoglobin in aqueous solution between 120 and 300 K showed that the friction coefficient was shown to linearly increase as temperature increases up to 300 K, independent of the glass transition temperature. This tendency cannot be well explained only by the Stokes–Einstein law, at least at 0 °C and higher, because the viscosity of liquid water decreases as a function of temperature. Thus, this tendency must have a different origin. Notably, the Langevin-based time correlation function should be fitted with care because the real time correlation will likely deviate for longer t when a constant value of γ is used without considering the time dependence, as in the generalized Langevin equation. The range of 0–5 ps was used for fitting the time correlation function in Ref. [44] and a very short (2–6 fs) Δt was used for the numerical derivative at $t = 0$ in Refs [15,27].

The Langevin mode is a multidimensional version of harmonic Langevin oscillators, formulated as a natural extension of NMA to include solvent effects [46,47]. Considering the $f \times f$ friction matrix $\mathbf{\Gamma}$ in addition to Hessian $\mathbf{F}$, the $f \times 2f$ eigenvector matrix $\mathbf{S}$ and the $2f \times 2f$ diagonal eigenvalue matrix $\boldsymbol{\xi} = \{\xi_\alpha\}$ are determined by the relations:

$$\mathbf{FS} + \mathbf{\Gamma S}\boldsymbol{\xi} + \mathbf{S}\boldsymbol{\xi}^2 = 0, \tag{25}$$

$$\mathbf{S}^T\mathbf{\Gamma S} + \boldsymbol{\xi}\mathbf{S}^T\mathbf{S} + \mathbf{S}^T\mathbf{S}\boldsymbol{\xi} = \boldsymbol{\xi}, \tag{26}$$

where the matrix elements of $\mathbf{L}$ and $\boldsymbol{\xi}$ are complex numbers. The friction matrix $\mathbf{\Gamma}$ can be modeled by diffusion tensors derived from hydrodynamics of polymers in solution [48–50]. Equations (25) and (26) correspond to (14) and (15) in NMA. It should be noted that the eigenvalue of the αth Langevin mode ξ_α can have both real and imaginary parts, which determine the damping factor and oscillatory frequency, respectively. In this case, the mode is underdamping. If ξ_α is a real number, the mode is overdamping. In the limit $\mathbf{\Gamma} \to 0$, the equations for NMA are recovered as:

$$\mathbf{S} = \frac{1}{\sqrt{2}}(\mathbf{W}, \mathbf{W}), \tag{27}$$

$$\boldsymbol{\xi} = \begin{pmatrix} \pm i\boldsymbol{\omega} & O \\ O & \pm i\boldsymbol{\omega} \end{pmatrix}. \tag{28}$$

A more general formulation based on the generalized Langevin equation and 3D reference interaction site model (3D-RISM) was proposed, called the Kim–Hirata theory [51,52]. This theory introduces the variance–covariance matrix obtained from the equilibrium free energy surface in solution determined by 3D-RISM and this matrix describes the force restoring an equilibrium conformation. Ref. [52] also proposed a protocol to evaluate friction based on the friction of an imaginary atom in solution from the site–site mode coupling theory [53,54], multiplied by the fraction of a protein atom contacting the solvent defined by the radial distribution function of solvent around the protein atom. The Kim–Hirata theory was further extended to analyze the temperature-dependent mean-square deviation of proteins [55].

6. Choice of Variables and Spaces for Better Representation of Macromolecular Dynamics in PCA

To successfully analyze macromolecular dynamics, a set of variables that well represent important features of dynamics should be employed for PCA. The Cartesian coordinates of atoms are frequently used. For a set of selected N atoms of the molecules of interest, the number of variables is $f = 3N$. The use of all-atom coordinates, including hydrogens, is useful for characterizing the anharmonic nature of protein dynamics directly compared to NMA [12,15,27,37,56]. The selection of C_α atoms is useful for selecting a small number

of large-amplitude motions, namely, "essential dynamics" [14]. Raw atomic coordinates from the original dataset typically reflect internal movements of the selected atoms, as well as overall translation and rotation. The translational and rotational components can be eliminated by the best fit of each dataset (typically a snapshot of a simulation trajectory) to a reference dataset so that the Eckart condition [57] is satisfied, for example, using the Kabsch method [58] or another method. To set the average $\langle \mathbf{q} \rangle$ as the origin of the coordinates, $\langle \mathbf{q} \rangle$ obtained after best fit should be used as the reference for the next round of best fit [12]. $\langle \mathbf{q} \rangle$ quickly converges within about five cycles with this procedure. Once translational and rotational components are completely eliminated, Cartesian PCA results in $(3N - 6)$ positive eigenvalues for internal motions and six zero eigenvalues corresponding to PCs of the translation and rotation.

Internal coordinates are also used frequently in PCA. In all-atom models of macromolecules, dihedral angles mainly determine the overall conformation of each molecule because the contributions of bond length and angle changes are relatively small. The significant movements of dihedral angles result in protein atoms moving non-linearly. Therefore, PCA in dihedral angle space can provide different information on protein dynamics compared to Cartesian PCA. If the deviation of dihedral angles from the average $\Delta\boldsymbol{\theta}$ is directly used as $\Delta\mathbf{q}$, the deviation of atomic coordinates from the average $\Delta\mathbf{r}$ is related to $\Delta\boldsymbol{\theta}$ as a first-order approximation as:

$$\Delta\mathbf{r} = \mathbf{L}\Delta\boldsymbol{\theta}, \tag{29}$$

where $\mathbf{L} = d\mathbf{r}/d\boldsymbol{\theta}$ is the Jacobian matrix that is evaluated for the average structure [59,60]. Omori et al. investigated the relation between the motions of atoms and dihedral angles and showed that the latter mutually move in a compensative manner, called "latent dynamics" [60]. The variance–covariance matrix of dihedral angles $\mathbf{C}_{\boldsymbol{\theta}}$ was compared to $\tilde{\mathbf{C}}_{\boldsymbol{\theta}}$ deduced from the atomic variance–covariance matrix $\mathbf{C}_{\mathbf{r}}$ as:

$$\tilde{\mathbf{C}}_{\boldsymbol{\theta}} = \left(\mathbf{L}^T \mathbf{C}_{\mathbf{r}}^{-1} \mathbf{L}\right)^{-1}. \tag{30}$$

Using backbone atoms (N, C_α and C) and dihedral angles (ϕ, ψ, and ω) of small globular proteins, Omori et al. also showed that $\tilde{\mathbf{C}}_{\boldsymbol{\theta}}$ precisely recovers the information of $\mathbf{C}_{\mathbf{r}}$ and contains higher-order dihedral correlations, but $\mathbf{C}_{\boldsymbol{\theta}}$ does not [60]. Additionally, the mean-square atomic displacements tended to be minimized upon rotation of the dihedral angles, indicating the compensative nature of dihedral dynamics. However, such latent dynamics behavior was not seen in dihedral PCs of deca-alanine, a short peptide.

The non-Euclidean nature of dihedral angles is not sufficiently considered in a linear transformation. PCA with dihedral angles may also require careful treatment as they are singular between 180° and −180°, also called a periodic boundary. Stock and coworkers proposed using dihedral angles differently in their dihedral angle PCA (dPCA), which uses cosines and sines of dihedral angle θ_l [61–63]:

$$\mathbf{q} = \{q_{2l-1}, q_{2l}\} = \{\cos\theta_l, \sin\theta_l\}, \tag{31}$$

where l denotes the dihedral angle index. This treatment can be considered as using the real and imaginary parts of $\exp i\theta_l$. Originally, Stock and coworkers focused on main chain ϕ and ψ [62], but it is straightforward to include other dihedral angles in this framework. In the case of short peptides, more rugged free energy landscapes were observed in the space spanned by the first few PCs in dPCA compared to the results of Cartesian PCA [61–63]. Based on the analysis of folding dynamics of the villin headpiece 35 (HP35), a small protein and native dynamics of BPTI up to the millisecond time scale, they reported that Cartesian PCA failed to capture important features of the free energy landscape and that dPCA gave a better presentation of the landscape [64].

Instead of projecting straight lines in the extrinsic tangent space of a mean, PCA for Riemannian manifolds was proposed based on geodesics of the intrinsic metric [65].

GeoPCA is a tool for dihedral angle-based principal component geodesics, in which angular data are projected on a sphere composed of the first two principal component geodesics [66]. GeoPCA was validated by using it to cluster a set of RNA conformations derived from a database comprising 73 RNA structures. Dihedral principal geodesic analysis (dPGA) was applied to reduce the dimension of the protein structure ensemble and the result was compared to the results obtained using PCA and dPCA [67].

The n-dimensional torus is a product space of n circles and can be used to characterize dihedral dynamics. Torus-PCA (T-PCA) was proposed and applied to the RNA benchmark used in GeoPCA [66], demonstrating the validity of T-PCA [68]. Another approach, dPCA+, minimizes residual projection error by transforming the data such that the maximal gap of the sampling is shifted to the periodic boundary of a dihedral angle [69]. Interestingly, this transformation also minimized the error of the covariance matrix. dPCA+ was also used to examine the non-equilibrium process simulated by targeted MD (TMD), and the free energy profiles of deca-alanine obtained by unbiased MD, Jarzynski identity and second-order cumulant approximation were compared [70]. In addition, the landscapes from unbiased MD, TMD and reweighted TMD were investigated in that report.

Other variables have been introduced to conduct PCA of simulation trajectories. To visualize MD and MC trajectories, Abagyan and Argos introduced a distance measure between two conformers, defined based on dihedral angles [71]. Java Essential Dynamics (JED) can use internal distance pair coordinates (dpPCA) as an option [72]. Ernst et al. introduced contact distance-based PCA (conPCA), reciprocal distance-based (iconPCA) and PCA based on inter-C_α distances (C_αPCA) and compared each result to those obtained using dPCA [73]. For conPCA, the distance between the closest heavy atom of each residue is considered as a contact if it is less than 4.5 Å and the residue pair of the contact is separated by more than three residues along the sequence [73,74]. Thus, conPCA can consider side chains in contact with each other but excludes information regarding local fluctuation along the sequence. Using 300 μs HP35 and 1 ms BPTI MD trajectories and examining the resolution of the free energy landscape and the decay of autocorrelation functions, Ernst et al. showed that distance-based PCAs, particularly C_αPCA, tend to be versatile, but they exhibit fewer landscape details than dPCA does [73]. Recently, Ogata proposed grid-based PCA (GBPCA), which considers a grid system consisting of cubes with 5 Å edges [75]. This method uses a unit vector of mass-weighted averages of atoms in each cube to calculate the correlations to be diagonalized and was applied to bulk water, bulk methane and hydrated proteins.

This review focuses on the PCA of structural data **Q**, but it is worth mentioning that PCA is also used for analyzing multiple spectra measured under different conditions [76]. Consider a set of spectra obtained as a function of frequency or wavelength measured under different temperatures, pH, times, etc. **Q** can be constructed with $i = 1, \ldots, f$ for frequency or wavelength and $m = 1, \ldots, M$ for the conditions. For example, rapid scanning wavelength stopped-flow kinetics experiments on liver alcohol dehydrogenase (LADH) whose reaction is spectrally complex were analyzed by PCA and absorbing species were identified [77,78]. Yuan et al. analyzed Fourier transform near-infrared (FT-NIR) spectra of bovine serum albumin (BSA) at temperatures ranging 45–85 °C by PCA and evolving factor analysis (EFA) [79]. The contributions from the first PCs obtained for two frequency ranges were both greater than 99%, indicating that most of the spectral variations are explained by the first PCs. PCA and EFA also revealed temperatures of structure changes. Sakurai and Goto conducted the PCA of pH dependence of heteronuclear sequential quantum correlation (HSQC) spectra of β-lactoglobulin measure by NMR spectroscopy and identified three conformational transitions at different pH [80]. These results validate the application of PCA for characterizing various condition-dependent spectra and for investigating structure changes, which also enables the PCA under a combination of multiple conditions, such as temperature-dependent kinetics.

As shown in Section 2, PCA and SVD provide the same information, **V**, **U** and **λ** from **Q**, but SVD directly determines these matrices without calculating **C**. Thus, SVD

is also employed for analyzing spectra similar to PCA. In spectral analysis, **U** quantifies condition-dependent components while **V** describes the condition-independent basis sets. By focusing on dominant components in **λ**, SVD can act as a mechanism-independent noise filter for spectra. Thus, SVD is regarded as an automated procedure of modeling of spectroscopic datasets [81,82] and is also used for the analysis of protein dynamics. For example, Hofrichter et al. measured time-dependent optical absorption spectra from 3 ns to 100 ms after the photolysis of the CO complex of hemoglobin and identified three significant basis spectra and five exponential relaxations from the time course of their amplitudes by SVD, which enabled the analysis of ligand rebinding kinetics [83]. Moffat and coworkers developed a method to analyze time-dependent difference electron density by SVD and analyzed structural intermediates of photoactive yellow protein (PYP) [84–86]. These works indicate the usefulness of SVD in spectral analysis.

7. The Fluctuation–Dissipation Theorem, Linear Response Theory and PCA

The fluctuation–dissipation theorem states that the linear response of a given system to an external perturbation is expressed in terms of fluctuation properties of the system in thermal equilibrium [87,88]. In a time-independent form, the linear response theory (LRT) shows that a perturbation applied to a system **f** results in response $\Delta\mathbf{q}_R$ mediated by the variance–covariance matrix **C** as:

$$\Delta\mathbf{q}_R \propto \mathbf{C}\mathbf{f}. \tag{32}$$

Using **C** obtained from MD simulations of an unliganded protein and **f** mimicking the protein–ligand interaction, LRT was shown to reproduce the response of the liganded protein [89]. Additionally, dihedral LRT based on the variance–covariance derived using Equation (30) was shown to better predict the ligand-bound form of ferric-binding protein [59]. Time-independent and time-dependent LRT showed agreement for the time response of myoglobin upon CO binding between LRT, ultraviolet resonance Raman spectroscopy and time-resolved X-ray crystallography, suggesting that the primary response can be described by LRT [90]. Hirata proposed a theory to evaluate a response function based on the aforementioned Kim–Hirata theory [91].

If LRT is considered in the PC space, the expected response (Equation (32)) in PC space $\boldsymbol{\sigma}_R$ is obtained in a manner similar to Equation (6) as:

$$\boldsymbol{\sigma}_R \propto \mathbf{V}^T\Delta\mathbf{q}_R. \tag{33}$$

Additionally, considering Equations (4) and (32), we obtain:

$$\boldsymbol{\sigma}_R \propto \boldsymbol{\lambda}\mathbf{V}^T\mathbf{f} = \boldsymbol{\lambda}\mathbf{f}_{PC}. \tag{34}$$

If **f** is applied as an isotropic random perturbation, the perturbation in PC space $\mathbf{f}_{PC} = \mathbf{V}^T\mathbf{f}$ is also isotropic. However, the response $\boldsymbol{\sigma}_R$ is scaled by **λ**, meaning that the perturbation force acting along the PC is proportional to λ_α [92]. Equation (34) also indicates that random perturbations are expected to cause highly anisotropic responses in the protein because proteins fluctuate in a highly anisotropic manner in equilibrium. Transform and relax sampling (TRS) enhances anisotropic protein movements, implicitly expecting the response in Equation (34) but without actually calculating **C** [92]. TRS is carried out as cycles of transform, relax and sampling stages. In the transform stage, the protein is perturbed by random forces during MD, then the protein is relaxed during MD without perturbation in the relax stage and finally usual MD is conducted as the sampling stage. TRS successfully simulated open-close motions of domains of multi-domain proteins several times within a simulation time of 20 ns. Additionally, folding–unfolding transitions of the "mini-protein" chignolin were observed many times during a 100 ns simulation.

Linear response path following (LRPF) simulates global conformational changes in proteins upon ligand binding by periodically updating a linear response (LR) force with three phases of MD, enabling non-linear transformation to a target direction [93]. In the

first phase, the LR force is obtained by computing **C** and a mean force acting from ligands during equilibrium MD. Biased MD subsequently induces conformational change in the second phase, then the final MD re-equilibrates the system without bias. LRPF predicted an inward-facing form of mitochondrial ADP/ATP carrier (AAC), a membrane transporter, starting from an outward-facing form determined by X-ray crystallography [94].

8. Non-Gaussianity and Non-Linearity in PCA

PCA performs best if the distribution of a dataset is a multidimensional Gaussian that depends only on the mean and variance–covariance (second moments), which are the only quantities considered to determine collective variables in PCA. However, a small number of dominant PCs show non-Gaussian distributions in protein dynamics. These non-Gaussian collective variables are believed to be important for protein function [12–15] and indicate a limitation of using second moments in determining new collective variables. Since non-Gaussian distributions cannot be well characterized with mean and second moments only, one solution may be to consider higher-order moments to determine collective variables other than PC. Independent component analysis (ICA) separates non-Gaussian signals that are independent from Gaussian noise and typically considers mutual information (MI) to quantify the correlations and determine the optimal coordinate transformation to minimize the MI [95–97]. Typical ICA employs preprocessing, "centering" and "whitening" [97]. "Centering" corresponds to the treatment already completed in PCA, as in Equation (1), in which the mean is subtracted. "Whitening" is typically the normalization of components by their standard deviations. Early applications of ICA for protein dynamics by full component analysis (FCA) used Cartesian coordinates [98] while the dihedral of Equation (31) was used for ICA in Ref. [99].

Most of the methods described in this review focus on extracting mutually independent components as much as possible, whereas independent subspace (ISA) determines significantly correlated collective motions. ISA features non-Gaussian behaviors similar to ICA, using fourth-order cumulants [100]. In this method, ISA based on the subspace joint approximate diagonalization of eigenmatrices algorithm (SJADE) [101] extracts several independent subspaces, in each of which collective modes are significantly correlated while the other modes are independent. Application of this method successfully detected the modes with long-tailed non-Gaussian probability distributions [100].

Another limitation is the linear transformation of PCA, whereas protein dynamics can be highly non-linear in nature. This problem was partially discussed in Section 6, mainly in relation to the use of dihedral angles in PCA and thus other PCA variants are discussed here. For example, Nguyen proposed the use of non-linear PCA (NLPCA) [102], enabled by non-linear mapping based on neutral networks [103].

Another possible solution for incorporating non-linearity in PCA is the introduction of kernel methods [104]. In kernel PCA [104], a new "feature space" F, which is non-linearly related to the original space, is introduced and principal components in F are considered. $\Delta\mathbf{q}_m$ is non-linearly mapped to a function $\mathbf{\Phi}(\Delta\mathbf{q}_m)$ that satisfies the condition $\langle\mathbf{\Phi}(\Delta\mathbf{q}_m)\rangle = 0$. Consider the variance–covariance in F as:

$$\mathbf{C} = \frac{1}{M}\sum_{m=1}^{M}\mathbf{\Phi}(\Delta\mathbf{q}_m)\,\mathbf{\Phi}^T(\Delta\mathbf{q}_m) = \left\langle\mathbf{\Phi}\,\mathbf{\Phi}^T\right\rangle, \tag{35}$$

and the eigenvalue problem shown in Equation (4). The nth eigenvector $\mathbf{v}_n$ is defined as a linear combination of $\mathbf{\Phi}(\Delta\mathbf{q}_m)$ with a coefficient matrix $\boldsymbol{\alpha} = \{\alpha_{mn}\}$ as:

$$\mathbf{v}_n = \sum_{m=1}^{M} \alpha_{mn} \mathbf{\Phi}(\Delta \mathbf{q}_m) \ . \tag{36}$$

Here, we introduce the kernel representation $k(\mathbf{x}, \mathbf{y}) = \mathbf{\Phi}(\mathbf{x})\mathbf{\Phi}^T(\mathbf{y})$ and an $M \times M$ matrix $\mathbf{K}$ as:

$$\mathbf{K} = \{k_{mn}\} = \{k(\Delta \mathbf{q}_m, \Delta \mathbf{q}_n)\}. \tag{37}$$

As shown in Ref. [104], $\boldsymbol{\alpha}$ is obtained by the relation:

$$\mathbf{K}\boldsymbol{\alpha} = M \, \boldsymbol{\lambda}\boldsymbol{\alpha}, \tag{38}$$

which can be solved by diagonalizing $\mathbf{K}$ using the condition:

$$\boldsymbol{\lambda}\boldsymbol{\alpha}^T\boldsymbol{\alpha} = \mathbf{I}. \tag{39}$$

Using $\mathbf{K}$ and $\boldsymbol{\alpha}$, principal components in F space are obtained as the projection of $\mathbf{\Phi}(\Delta \mathbf{q})$ onto the nth eigenvector by:

$$\mathbf{v}_n^T \mathbf{\Phi}(\Delta \mathbf{q}_m) = \frac{1}{\lambda_n} \sum_{m=1}^{M} \alpha_{mn} k_{mn} \ . \tag{40}$$

Equations (38) and (40) indicate that eigenvalues and principal components in F space are determined from the kernel $\mathbf{K}$ without directly solving Equation (35), which is typically difficult. It is worth mentioning that the use of $k(\mathbf{x}, \mathbf{y}) = \mathbf{x}\mathbf{y}^T$ recovers the original PCA and the use of $k(\mathbf{x}, \mathbf{y}) = \left(\mathbf{x}\mathbf{y}^T\right)^d$ $(d > 1)$ considers higher moments. Additionally, Gaussian kernel $k(\mathbf{x}, \mathbf{y}) = \exp\left(-\frac{\|\mathbf{x}-\mathbf{y}\|^2}{2\sigma^2}\right)$ with the adjustable parameter σ is frequently used in kernel methods. Kernel PCA can also be applied to the analysis of protein dynamics [21].

"Diffusion maps" is used as nonlinear dimensionality reduction method to embed data points into a Euclidean space in which the Euclidean distance is equal to "diffusion distance" and to conduct the reduction by neglecting certain dimensions in the diffusion space [105–107]. Diffusion maps considers a Markov chain random walk on normalized data points and the probability of one-step random walk between two data points (connectivity), which is proportional a kernel function (typically Gaussian kernel). For diffusion maps, "diffusion matrix" is obtained by normalizing the rows of the kernel matrix, eigenvectors of the diffusion matrix are calculated and the diffusion mapping is conducted by mapping the data points to dominant eigenvectors in the diffusion space. Ferguson et al. used diffusion maps for a protein simulation for the first time to investigate trajectories of replica-exchange molecular dynamics (REMD) simulations of pro-microcin J25 (pro-MccJ25), the 21-residue uncyclized analog of antimicrobial peptide microcin J25 (MccJ25), which identified two global order parameters and three distinct pathways of conformational change [108]. One of the pathways was shown to correspond to a conformational change to left-handed lasso coil conformations. Kim et al. employed diffusion maps to characterize MD trajectories of Trp-cage miniprotein folding [109]. Recently, Trstanova et al. demonstrated the use of diffusion maps to identify metastable states as well as to formalize the locality within the metastable states in the analysis of molecular systems [110].

9. Detecting Data Differences by PCA and Related Methods

PCA is also used to detect differences in a dataset or featurize the differences between datasets. Self-consistency of dipolar couplings analysis (SECONDA) is a PCA based on residual dipolar coupling (RDS) measured by NMR and was developed to examine the effects of different alignment media used for RDS measurements [111]. The results showed that if the structure and dynamics of the target molecule are the same, PCA gives at most five nonzero eigenvalues, but additional nonzero eigenvalues are obtained if differences exist.

Howe conducted Cartesian PCA of 49 NMR-determined structures of EF40, a 28-residue peptide and demonstrated that the structures are clustered and outliers are detected [112]. Using the online tool PCA_NEST, systematic analysis of 24 pairs of enzyme structures determined by both solution NMR and X-ray crystallography revealed differences in the solution and crystal structures of the proteins [113]. Notably, the X-ray structures were shown to be a conformational state along the dominant PCs derived from NMR models, consistent with the expectation from LRT, since the environmental differences (in crystal or in solution) can be considered as a perturbation to the protein.

Sakuraba and Kono employed linear discriminant analysis with iterative procedure (LDA-ITER) to compare two trajectories obtained under different conditions [114]. LDA-ITER was developed to consider the trace ratio optimization problem in supervised learning and maximizes the ratio of two matrices by unitary transformation [115,116]. This method finds the axis of projection that separates two trajectories while keeping each trajectory well clustered. LDA-ITER was applied to the wild-type and R96H mutant of T4 lysozyme, as well as to the liganded and unliganded PDZ2 domains of human phosphatase hPTP1E. The results showed very clear separation of two kinds of trajectories along the first dimension and a Gaussian-like distribution for each cluster. In contrast, the projections onto the first two PCAs significantly overlapped, and another method, partial least squares discrimination analysis (PLS-DA) [117], gave less overlapped results in the 2D projection compared to PCA. However, LDA-ITER gave the best results. In addition, important differences were characterized on a residue-by-residue basis.

Relative principal component analysis (RPCA) also featurizes the differences between two datasets, but the transformation is determined by maximizing the Kullback−Leibler divergence between the probability distributions between the two states and by simultaneously diagonalizing two variance–covariance matrices of the states with a single transformation matrix [118]. The application of RPCA to HIV-1 protease showed better performance compared to PCA and identified conformational hotspots.

10. Time Evolution of Collective Variables

The methods described in Sections 4 and 5 consider the time evolution of collective variables based on physical models, but another type of approach investigates evolution from a phenological perspective. Time-independent component analysis (tICA) is a variation of ICA that focuses on the time independence between time 0 and τ, instead of considering non-Gaussianity [119–121]. In addition to $\mathbf{C}$ defined in Equation (1), tICA introduces the time-lagged variance–covariance matrix:

$$\mathbf{C}(\tau) = \left\langle \Delta\mathbf{q}\,(0)\Delta\mathbf{q}^T(\tau) \right\rangle, \tag{41}$$

using the time lag parameter τ. Although $\mathbf{C} = \mathbf{C}(0)$ is symmetric, $\mathbf{C}(\tau)$ can be a non-symmetric matrix. In tICA, the generalized eigenvalue problem is solved with the normalization condition:

$$\mathbf{C}(\tau)\mathbf{Y} = \mathbf{C}\mathbf{Y}\zeta, \tag{42}$$

$$\mathbf{Y}^T\mathbf{C}\mathbf{Y} = \mathbf{I}, \tag{43}$$

where $\zeta = \{\zeta_\alpha\}$ and $\mathbf{Y} = \left(\mathbf{y}_1 \cdots \mathbf{y}_f\right)$ are eigenvalue and eigenvector matrices, respectively. Since matrix elements of ζ and $\mathbf{Y}$ are generally obtained as complex numbers, the symmetrization $\bar{\mathbf{C}}(\tau) = \frac{1}{2}\left(\mathbf{C}(\tau) + \mathbf{C}^T(\tau)\right)$ conducted in Refs [120,121] and Equation (42) is modified as:

$$\bar{\mathbf{C}}(\tau)\mathbf{Y} = \mathbf{C}\mathbf{Y}\zeta, \tag{44}$$

such that ζ and $\mathbf{Y}$ comprise real values. Combining Equation (42) or (44) with (43) gives:

$$\mathbf{Y}^T\mathbf{C}(\tau)\mathbf{Y} = \zeta \text{ or} \mathbf{Y}^T\bar{\mathbf{C}}(\tau)\mathbf{Y} = \zeta, \tag{45}$$

respectively. Equations (43) and (45) indicate that the autocorrelation function of the αth time-independent component at time 0 is $\mathbf{y}_\alpha^T\mathbf{C}\mathbf{y}_\alpha = 1$ and that at time τ is $\mathbf{y}_\alpha^T\bar{\mathbf{C}}(\tau)\mathbf{y}_\alpha = \zeta_\alpha$. If the autocorrelation function is a single exponential decay with characteristic time T_α, we obtain the relation $\exp(-\tau/T_\alpha) = \zeta_\alpha$, which results in:

$$T_\alpha = -\frac{\tau}{\ln \zeta_\alpha}. \tag{46}$$

Dynamic component analysis (DCA) considers the time-lagged variance–covariance matrix for normalized PCs, but Equation (41) is recovered in the original coordinates [122], indicating the equivalence of tICA and DCA in this formulation. In practice, however, DCA uses the inter-residue distance (distance map) as the coordinates of PCA [122].

Relaxation mode analysis (RMA) [123–127] considers "whitened" signals as the αth relaxation mode $\phi_\alpha(t)$ as:

$$\mathbf{c}(\tau) = \{\langle \phi_\alpha(0)\phi_\beta(t)\rangle\} = \{\delta_{\alpha\beta}\exp(-\kappa_\alpha\tau)\}, \tag{47}$$

where $\boldsymbol{\kappa} = \{\kappa_\alpha\}$ and $\delta_{\alpha\beta}$ indicate the relaxation rate and Kroneker delta, respectively. Independence and single exponential decay of each relaxation mode are assumed in Equation (47). Using $\mathbf{c}(\tau)$ values at time τ and $\mathbf{c} = \mathbf{c}(0)$, the relaxation rates are determined as:

$$\mathbf{c}(\tau)\mathbf{y} = \exp(-\boldsymbol{\kappa}\tau)\mathbf{c}\mathbf{y}, \tag{48}$$

$$\mathbf{y}^T\mathbf{c}\mathbf{y} = \mathbf{I}. \tag{49}$$

These equations correspond to Equations (42) and (43) for tICA, indicating the equivalence of RMA and tICA, and that RMA in practice can be considered as tICA for the whitened signals. Similar to tICA, symmetrization $\bar{\mathbf{c}}(\tau) = \frac{1}{2}(\mathbf{c}(\tau) + \mathbf{c}^T(\tau))$ is also used in RMA. Both tICA and RMA can obtain T_α or κ_α, which characterize a single exponential decay.

In principle, the results of these methods depend on the time lag parameter τ. The rigid-body domain motion of lysine-, arginine- and ornithine-binding protein (LAO) was analyzed by tICA in the ligand-free open state during a 600 ns MD using a 10 ns lag time, and the IC vectors and relaxation rates were shown to be fairly robust on this time scale [120]. tICA of LAO backbone dynamics for 1 μs was conducted using a 1 ns lag time [121], whereas DMA of the folding/unfolding dynamics of the FiP35 WW domain for two 100 μs MD trajectories used a 10 ns lag time. RMA of a 10-residue peptide, chignolin, for a 750 ns MD simulation at 450 K in solution used a relatively short lag time (10 ps) [127]. A two-step RMA was proposed and used to conduct a 2 μs MD trajectory of hen egg-white lysozyme [128].

tICA projections of high-dimensional random walks were recently shown to resemble cosine functions, be strongly dependent on the lag time and be very similar to those of 1 μs MD trajectories of proteins, particularly for larger proteins [129]. Although the introduction of a lag time allows tICA to provide richer information than PCA, care must be taken in choosing the lag time.

Time-dependent PCA (TDPCA), proposed recently, conducts multiple PCAs for short segments taken from a single MD trajectory by shifting the time window and allowing overlap, which provides time-dependent eigenvalues and eigenvectors. This approach was applied to a bulk water model and a coarse-grained protein-G model [130].

tICA and PCA are widely used as dimension reduction methods in the Markov state model (MSM) [131–133]. Notably, kernel methods were combined with tICA to provide kernel tICA (ktICA) for MSM [134]. The use of tICA-related methods in MSM is described in detail in recent review articles on MSM [135,136].

11. Conclusions

This review reported the development of PCA and related methods for analyzing protein dynamics, from basic concepts to the latest advanced methods. Possible applications of PCA are now very broad and many variations in PCA have been developed for specific purposes. As is clear from this review, it is difficult to specify the best and most versatile PCA method. Rather, the most suitable method should be chosen based on the purpose of the simulation and analysis. In many of the examples described, improved methods provided better performance than "classical" PCA. Additionally, the choice of original coordinates is important, as shown in Section 6. In addition, different methods should be tested, allowing the best method to be selected or to examine the validity of the obtained conclusion.

It is worth mentioning that some of the advanced methods described in this review do not directly employ original coordinates for large molecules but rather use a two-step procedure. Specifically, standard PCA and dimensional reduction into a smaller subspace are conducted first, then more advanced component analysis is conducted. For example, FCA [98] and ISA [101] described in these references used 100 PCs. DCA described in [122] used dimensional reduction by PCA but the number of PCs employed is unclear to us. The top five PCs were used in kPCA [21]. Therefore, even in cases where more sophisticated methods than standard PCA are used, PCA can be used for dimensionality reduction and as a reference for the comparison of other PCA-related methods.

Author Contributions: Conceptualization, A.K.; investigation, A.K.; writing—original draft preparation, review and editing, A.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by MEXT/JSPS KAKENHI Nos. JP19H03191, JP20H05439, JP21H05510, and JP22H04745, and by MEXT as a "Program for Promoting Researches on the Supercomputer Fugaku" (Application of Molecular Dynamics Simulation to Precision Medicine Using Big Data Integration System for Drug Discovery, JPMXP1020200201, and Biomolecular Dynamics in a Living Cell, JPMXP1020200101) to A.K.

Conflicts of Interest: The author declares no conflict of interest.

References

1. Pearson, K.L., III. On lines and planes of closest fit to systems of points in space. *Lond. Edinb. Dublin Philos. Mag. J. Sci.* **1901**, *2*, 559–572. [CrossRef]
2. Hotelling, H. Analysis of a complex of statistical variables into principal components. *J. Educ. Psychol.* **1933**, *24*, 417–441. [CrossRef]
3. Mccammon, J.A.; Gelin, B.R.; Karplus, M. Dynamics of Folded Proteins. *Nature* **1977**, *267*, 585–590. [CrossRef] [PubMed]
4. Go, N.; Noguti, T.; Nishikawa, T. Dynamics of a Small Globular Protein in Terms of Low-Frequency Vibrational-Modes. *Proc. Natl. Acad. Sci. USA* **1983**, *80*, 3696–3700. [CrossRef] [PubMed]
5. Levitt, M.; Sander, C.; Stern, P.S. The normal modes of a protein: Native bovine pancreatic trypsin inhibitor. *Int. J. Quant. Chem.* **1983**, *24*, 181–199. [CrossRef]
6. Brooks, B.; Karplus, M. Harmonic dynamics of proteins: Normal modes and fluctuations in bovine pancreatic trypsin inhibitor. *Proc. Natl. Acad. Sci. USA* **1983**, *80*, 6571–6575. [CrossRef]
7. Richards, F.M. The interpretation of protein structures: Total volume, group volume distributions and packing density. *J. Mol. Biol.* **1974**, *82*, 1–14. [CrossRef]
8. Karplus, M.; Kushick, J.N. Method for estimating the configurational entropy of macromolecules. *Macromolecules* **1981**, *14*, 325–332. [CrossRef]
9. Levy, R.M.; Srinivasan, A.R.; Olson, W.K.; McCammon, J.A. Quasi-harmonic method for studying very low frequency modes in proteins. *Biopolymers* **1984**, *23*, 1099–1112. [CrossRef]
10. Levy, R.M.; Rojas, O.D.; Friesner, R.A. Quasi-Harmonic Method for Calculating Vibrational-Spectra from Classical Simulations on Multidimensional Anharmonic Potential Surfaces. *J. Phys. Chem.* **1984**, *88*, 4233–4238. [CrossRef]
11. Horiuchi, T.; Go, N. Projection of Monte Carlo and molecular dynamics trajectories onto the normal mode axes: Human lysozyme. *Proteins Struct. Funct. Genet.* **1991**, *10*, 106–116. [CrossRef]
12. Kitao, A.; Hirata, F.; Gō, N. The effects of solvent on the conformation and the collective motions of protein: Normal mode analysis and molecular dynamics simulations of melittin in water and in vacuum. *Chem. Phys.* **1991**, *158*, 447–472. [CrossRef]
13. García, A.E. Large-Amplitude Nonlinear Motions in Proteins. *Phys. Rev. Lett.* **1992**, *68*, 2696–2699. [CrossRef] [PubMed]

14. Amadei, A.; Linssen, A.B.M.; Berendsen, H.J.C. Essential Dynamics of Proteins. *Proteins Struct. Funct. Genet.* **1993**, *17*, 412–425. [CrossRef] [PubMed]
15. Hayward, S.; Kitao, A.; Hirata, F.; Go, N. Effect of solvent on collective motions in globular protein. *J. Mol. Biol.* **1993**, *234*, 1207–1217. [CrossRef]
16. Maisuradze, G.G.; Liwo, A.; Scheraga, H.A. Principal component analysis for protein folding dynamics. *J. Mol. Biol.* **2009**, *385*, 312–329. [CrossRef]
17. Maisuradze, G.G.; Liwo, A.; Senet, P.; Scheraga, H.A. Local vs global motions in protein folding. *J. Chem. Theory Comput.* **2013**, *9*, 2907–2921. [CrossRef]
18. Hayward, S.; Go, N. Collective Variable Description of Native Protein Dynamics. *Annu. Rev. Phys. Chem.* **1995**, *46*, 223–250. [CrossRef]
19. Kitao, A.; Go, N. Investigating protein dynamics in collective coordinate space. *Curr. Opin. Struct. Biol.* **1999**, *9*, 164–169. [CrossRef]
20. Berendsen, H.J.C.; Hayward, S. Collective protein dynamics in relation to function. *Curr. Opin. Struct. Biol.* **2000**, *10*, 165–169. [CrossRef]
21. David, C.C.; Jacobs, D.J. Principal component analysis: A method for determining the essential dynamics of proteins. *Methods Mol. Biol.* **2014**, *1084*, 193–226. [CrossRef] [PubMed]
22. Kitao, A.; Takemura, K. High anisotropy and frustration: The keys to regulating protein function efficiently in crowded environments. *Curr. Opin. Struct. Biol.* **2017**, *42*, 50–58. [CrossRef] [PubMed]
23. Sittel, F.; Stock, G. Perspective: Identification of collective variables and metastable states of protein dynamics. *J. Chem. Phys.* **2018**, *149*, 150901. [CrossRef] [PubMed]
24. Davis, I.W.; Arendall, W.B., 3rd; Richardson, D.C.; Richardson, J.S. The backrub motion: How protein backbone shrugs when a sidechain dances. *Structure* **2006**, *14*, 265–274. [CrossRef]
25. Hayward, S. Peptide-plane flipping in proteins. *Protein Sci.* **2001**, *10*, 2219–2227. [CrossRef]
26. Nishima, W.; Qi, G.; Hayward, S.; Kitao, A. DTA: Dihedral transition analysis for characterization of the effects of large main-chain dihedral changes in proteins. *Bioinformatics* **2009**, *25*, 628–635. [CrossRef]
27. Kitao, A.; Hayward, S.; Go, N. Energy landscape of a native protein: Jumping-among-minima model. *Proteins Struct. Funct. Genet.* **1998**, *33*, 496–517. [CrossRef]
28. Joti, Y.; Kitao, A.; Go, N. Protein boson peak originated from hydration-related multiple minima energy landscape. *J. Am. Chem. Soc.* **2005**, *127*, 8705–8709. [CrossRef]
29. Kitao, A.; Wagner, G. A space-time structure determination of human CD2 reveals the CD58-binding mode. *Proc. Natl. Acad. Sci. USA* **2000**, *97*, 2064–2068. [CrossRef]
30. Kitao, A.; Wagner, G. Amplitudes and directions of internal protein motions from a JAM analysis of 15N relaxation data. *Magn. Reson. Chem.* **2006**, *44*, S130–S142. [CrossRef]
31. Hess, B. Similarities between principal components of protein dynamics and random diffusion. *Phys. Rev. E* **2000**, *62*, 8438–8448. [CrossRef] [PubMed]
32. Edelman, A.; Rao, N.R. Random matrix theory. *Acta Numer.* **2005**, *14*, 233–297. [CrossRef]
33. Kwapień, J.; Drożdż, S. Physical approach to complex systems. *Phys. Rep.* **2012**, *515*, 115–226. [CrossRef]
34. Palese, L.L. Random Matrix Theory in molecular dynamics analysis. *Biophys. Chem.* **2015**, *196*, 1–9. [CrossRef] [PubMed]
35. Palese, L.L. A random version of principal component analysis in data clustering. *Comput. Biol. Chem.* **2018**, *73*, 57–64. [CrossRef] [PubMed]
36. Cossio-Perez, R.; Palma, J.; Pierdominici-Sottile, G. Consistent Principal Component Modes from Molecular Dynamics Simulations of Proteins. *J. Chem. Inf. Model.* **2017**, *57*, 826–834. [CrossRef] [PubMed]
37. Hayward, S.; Kitao, A.; Go, N. Harmonicity and anharmonicity in protein dynamics: A normal mode analysis and principal component analysis. *Proteins Struct. Funct. Genet.* **1995**, *23*, 177–186. [CrossRef]
38. Bahar, I.; Rader, A.J. Coarse-grained normal mode analysis in structural biology. *Curr. Opin. Struct. Biol.* **2005**, *15*, 586–592. [CrossRef] [PubMed]
39. Ma, J. Usefulness and limitations of normal mode analysis in modeling dynamics of biomolecular complexes. *Structure* **2005**, *13*, 373–380. [CrossRef]
40. Normal Mode Analysis: Theory and Applications to Biological and Chemical Systems; Cui, Q. , *Bahar, I., Eds.*; Chapman & Hall/CRC: Boca Raton, FL, USA, 2006.
41. Dykeman, E.C.; Sankey, O.F. Normal mode analysis and applications in biological physics. *J. Phys. Condens. Matter* **2010**, *22*, 423202. [CrossRef]
42. Yamato, T.; Laprevote, O. Normal mode analysis and beyond. *Biophys. Physicobiol.* **2019**, *16*, 322–327. [CrossRef]
43. Jacob, A.B.; Vladena, B.-H. Normal Mode Analysis: A Tool for Better Understanding Protein Flexibility and Dynamics with Application to Homology Models. In *Homology Molecular Modeling*; Rafael Trindade, M., de Moraes, F.R.M., Magnólia, C., Eds.; IntechOpen: Rijeka, Croatia, 2021.

44. Moritsugu, K.; Smith, J.C. Langevin model of the temperature and hydration dependence of protein vibrational dynamics. *J. Phys. Chem. B* **2005**, *109*, 12182–12194. [CrossRef]
45. Moritsugu, K.; Smith, J.C. Temperature-dependent protein dynamics: A simulation-based probabilistic diffusion-vibration Langevin description. *J. Phys. Chem. B* **2006**, *110*, 5807–5816. [CrossRef]
46. Lamm, G.; Szabo, A. Langevin Modes of Macromolecules. *J. Chem. Phys.* **1986**, *85*, 7334–7348. [CrossRef]
47. Kottalam, J.; Case, D.A. Langevin Modes of Macromolecules—Applications to Crambin and DNA Hexamers. *Biopolymers* **1990**, *29*, 1409–1421. [CrossRef]
48. Kirkwood, J.G.; Riseman, J. The Intrinsic Viscosities and Diffusion Constants of Flexible Macromolecules in Solution. *J. Chem. Phys.* **1948**, *16*, 565–573. [CrossRef]
49. Kirkwood, J.G. The statistical mechanical theory of irreversible processes in solutions of flexible macromolecules. Visco-elastic behavior. *Recl. Trav. Chim. Pays-Bas* **1949**, *68*, 649–660. [CrossRef]
50. Rotne, J.; Prager, S. Variational Treatment of Hydrodynamic Interaction in Polymers. *J. Chem. Phys.* **1969**, *50*, 4831–4837. [CrossRef]
51. Kim, B.; Hirata, F. Structural fluctuation of protein in water around its native state: A new statistical mechanics formulation. *J. Chem. Phys.* **2013**, *138*, 054108. [CrossRef]
52. Hirata, F.; Kim, B. Multi-scale dynamics simulation of protein based on the generalized Langevin equation combined with 3D-RISM theory. *J. Mol. Liq.* **2016**, *217*, 23–28. [CrossRef]
53. Chong, S.-H.; Hirata, F. Dynamics of solvated ion in polar liquids: An interaction-site-model description. *J. Chem. Phys.* **1998**, *108*, 7339–7349. [CrossRef]
54. Chong, S.-H.; Hirata, F. Dynamics of ions in liquid water: An interaction-site-model description. *J. Chem. Phys.* **1999**, *111*, 3654–3667. [CrossRef]
55. Hirata, F. On the interpretation of the temperature dependence of the mean square displacement (MSD) of protein, obtained from the incoherent neutron scattering. *J. Mol. Liq.* **2018**, *270*, 218–226. [CrossRef]
56. Hayward, S.; Kitao, A.; Go, N. Harmonic and anharmonic aspects in the dynamics of BPTI: A normal mode analysis and principal component analysis. *Protein Sci.* **1994**, *3*, 936–943. [CrossRef] [PubMed]
57. Eckart, C. Some studies concerning rotating axes and polyatomic molecules. *Phys. Rev.* **1935**, *47*, 552–558. [CrossRef]
58. Kabsch, W. Solution for Best Rotation to Relate 2 Sets of Vectors. *Acta Crystallogr. A* **1976**, *32*, 922–923. [CrossRef]
59. Omori, S.; Fuchigami, S.; Ikeguchi, M.; Kidera, A. Linear response theory in dihedral angle space for protein structural change upon ligand binding. *J. Comput. Chem.* **2009**, *30*, 2602–2608. [CrossRef]
60. Omori, S.; Fuchigami, S.; Ikeguchi, M.; Kidera, A. Latent dynamics of a protein molecule observed in dihedral angle space. *J. Chem. Phys.* **2010**, *132*, 115103. [CrossRef]
61. Mu, Y.G.; Nguyen, P.H.; Stock, G. Energy landscape of a small peptide revealed by dihedral angle principal component analysis. *Proteins* **2005**, *58*, 45–52. [CrossRef]
62. Altis, A.; Nguyen, P.H.; Hegger, R.; Stock, G. Dihedral angle principal component analysis of molecular dynamics simulations. *J. Chem. Phys.* **2007**, *126*, 244111. [CrossRef] [PubMed]
63. Altis, A.; Otten, M.; Nguyen, P.H.; Hegger, R.; Stock, G. Construction of the free energy landscape of biomolecules via dihedral angle principal component analysis. *J. Chem. Phys.* **2008**, *128*, 245102. [CrossRef] [PubMed]
64. Sittel, F.; Jain, A.; Stock, G. Principal component analysis of molecular dynamics: On the use of Cartesian vs. internal coordinates. *J. Chem. Phys.* **2014**, *141*, 014111. [CrossRef] [PubMed]
65. Huckemann, S.; Ziezold, H. Principal component analysis for Riemannian manifolds, with an application to triangular shape spaces. *Adv. Appl. Probab.* **2006**, *38*, 299–319. [CrossRef]
66. Sargsyan, K.; Wright, J.; Lim, C. GeoPCA: A new tool for multivariate analysis of dihedral angles based on principal component geodesics. *Nucleic Acids Res.* **2012**, *40*, e25, Erratum in *Nucleic Acids Res.* **2015**, *43*, 10571–10572. [CrossRef] [PubMed]
67. Nodehi, A.; Golalizadeh, M.; Heydari, A. Dihedral angles principal geodesic analysis using nonlinear statistics. *J. Appl. Stat.* **2015**, *42*, 1962–1972. [CrossRef]
68. Eltzner, B.; Huckemann, S.; Mardia, K.V. Torus principal component analysis with applications to RNA structure. *J. Appl. Stat.* **2018**, *12*, 1332–1359. [CrossRef]
69. Sittel, F.; Filk, T.; Stock, G. Principal component analysis on a torus: Theory and application to protein dynamics. *J. Chem. Phys.* **2017**, *147*, 244101. [CrossRef]
70. Post, M.; Wolf, S.; Stock, G. Principal component analysis of nonequilibrium molecular dynamics simulations. *J. Chem. Phys.* **2019**, *150*, 204110. [CrossRef]
71. Abagyan, R.; Argos, P. Optimal protocol and trajectory visualization for conformational searches of peptides and proteins. *J. Mol. Biol.* **1992**, *225*, 519–532. [CrossRef]
72. David, C.C.; Singam, E.R.A.; Jacobs, D.J. JED: A Java Essential Dynamics Program for comparative analysis of protein trajectories. *BMC Bioinform.* **2017**, *18*, 271. [CrossRef]
73. Ernst, M.; Sittel, F.; Stock, G. Contact- and distance-based principal component analysis of protein dynamics. *J. Chem. Phys.* **2015**, *143*, 244114. [CrossRef] [PubMed]

74. Heringa, J.; Argos, P. Side-chain clusters in protein structures and their role in protein folding. *J. Mol. Biol.* **1991**, *220*, 151–171. [CrossRef]
75. Ogata, K. Investigation of Cooperative Modes for Collective Molecules Using Grid-Based Principal Component Analysis. *J. Phys. Chem. B* **2021**, *125*, 1072–1084. [CrossRef]
76. Beattie, J.R.; Esmonde-White, F.W.L. Exploration of Principal Component Analysis: Deriving Principal Component Analysis Visually Using Spectra. *Appl. Spectrosc.* **2021**, *75*, 361–375. [CrossRef] [PubMed]
77. Cochran, R.N.; Horne, F.H. Strategy for resolving rapid scanning wavelength experiments by principal component analysis. *J. Phys. Chem.* **1980**, *84*, 2561–2567. [CrossRef]
78. Cochran, R.N.; Horne, F.H.; Dye, J.L.; Ceraso, J.; Suelter, C.H. Principal component analysis of rapid scanning wavelength stopped-flow kinetics experiments on the liver alcohol dehydrogenase catalyzed reduction of p-nitroso-N,N-dimethylaniline by 1,4-dihydronicotinamide adenine dinucleotide. *J. Phys. Chem.* **1980**, *84*, 2567–2575. [CrossRef]
79. Yuan, B.; Murayama, K.; Wu, Y.; Tsenkova, R.; Dou, X.; Era, S.; Ozaki, Y. Temperature-dependent near-infrared spectra of bovine serum albumin in aqueous solutions: Spectral analysis by principal component analysis and evolving factor analysis. *Appl. Spectrosc.* **2003**, *57*, 1223–1229. [CrossRef]
80. Sakurai, K.; Goto, Y. Principal component analysis of the pH-dependent conformational transitions of bovine beta-lactoglobulin monitored by heteronuclear NMR. *Proc. Natl. Acad. Sci. USA* **2007**, *104*, 15346–15351. [CrossRef]
81. Henry, E.R. The Use of Matrix Methods in the Modeling of Spectroscopic Data Sets. *Biophys. J.* **1997**, *72*, 652–673. [CrossRef]
82. Shrager, R.I.; Hendler, R.W. Titration of individual components in a mixture with resolution of difference spectra, pKs, and redox transitions. *Anal. Chem.* **2002**, *54*, 1147–1152. [CrossRef]
83. Hofrichter, J.; Sommer, J.H.; Henry, E.R.; Eaton, W.A. Nanosecond absorption spectroscopy of hemoglobin: Elementary processes in kinetic cooperativity. *Proc. Natl. Acad. Sci. USA* **1983**, *80*, 2235–2239. [CrossRef]
84. Schmidt, M.; Rajagopal, S.; Ren, Z.; Moffat, K. Application of Singular Value Decomposition to the Analysis of Time-Resolved Macromolecular X-Ray Data. *Biophys. J.* **2003**, *84*, 2112–2129. [CrossRef]
85. Rajagopal, S.; Schmidt, M.; Anderson, S.; Ihee, H.; Moffat, K. Analysis of experimental time-resolved crystallographic data by singular value decomposition. *Acta Crystallogr. D* **2004**, *60*, 860–871. [CrossRef]
86. Kostov, K.S.; Moffat, K. Cluster analysis of time-dependent crystallographic data: Direct identification of time-independent structural intermediates. *Biophys. J.* **2011**, *100*, 440–449. [CrossRef]
87. Kubo, R. The fluctuation-dissipation theorem. *Rep. Prog. Phys.* **1966**, *29*, 255–284. [CrossRef]
88. Des Cloizeaux, D. Linear Response, Generalized Susceptibility and Dispersion Theory. In *Theory of Condensed Matter*; Bassani, F., Caglioti, G., Ziman, J., Eds.; International Center for Theretical Physics: Trieste, Italy, 1968; pp. 325–354.
89. Ikeguchi, M.; Ueno, J.; Sato, M.; Kidera, A. Protein structural change upon ligand binding: Linear response theory. *Phys. Rev. Lett.* **2005**, *94*, 078102. [CrossRef]
90. Yang, L.W.; Kitao, A.; Huang, B.C.; Go, N. Ligand-Induced Protein Responses and Mechanical Signal Propagation Described by Linear Response Theories. *Biophys. J.* **2014**, *107*, 1415–1425. [CrossRef]
91. Hirata, F. A molecular theory of the structural dynamics of protein induced by a perturbation. *J. Chem. Phys.* **2016**, *145*, 234106. [CrossRef]
92. Kitao, A. Transform and relax sampling for highly anisotropic systems: Application to protein domain motion and folding. *J. Chem. Phys.* **2011**, *135*, 045101, Erratum in *J. Chem. Phys.* **2011**, *135*, 119903. [CrossRef]
93. Tamura, K.; Hayashi, S. Linear Response Path Following: A Molecular Dynamics Method To Simulate Global Conformational Changes of Protein upon Ligand Binding. *J. Chem. Theory Comput.* **2015**, *11*, 2900–2917. [CrossRef]
94. Tamura, K.; Hayashi, S. Atomistic modeling of alternating access of a mitochondrial ADP/ATP membrane transporter with molecular simulations. *PLoS ONE* **2017**, *12*, e0181489. [CrossRef]
95. Jutten, C.; Herault, J. Blind separation of sources, part I: An adaptive algorithm based on neuromimetic architecture. *Signal Process.* **1991**, *24*, 1–10. [CrossRef]
96. Comon, P. Independent component analysis, A new concept? *Signal Process.* **1994**, *36*, 287–314. [CrossRef]
97. Hyvärinen, A.; Oja, E. Independent component analysis: Algorithms and applications. *Neural Netw.* **2000**, *13*, 411–430. [CrossRef]
98. Lange, O.F.; Grubmüller, H. Full correlation analysis of conformational protein dynamics. *Proteins* **2008**, *70*, 1294–1312. [CrossRef]
99. Nguyen, P.H. Conformational states and folding pathways of peptides revealed by principal-independent component analyses. *Proteins* **2007**, *67*, 579–592. [CrossRef]
100. Sakuraba, S.; Joti, Y.; Kitao, A. Detecting coupled collective motions in protein by independent subspace analysis. *J. Chem. Phys.* **2010**, *133*, 185102. [CrossRef]
101. Theis, F.J. Towards a general independent subspace analysis. In *Advances in Neural Information Processing Systems*; Schölkopf, B., Platt, J., Hoffman, T.E., Eds.; MIT Press: Cambridge, MA, USA, 2007; Volume 19, pp. 1361–1368.
102. Nguyen, P.H. Complexity of free energy landscapes of peptides revealed by nonlinear principal component analysis. *Proteins* **2006**, *65*, 898–913. [CrossRef]
103. Kramer, M.A. Nonlinear principal component analysis using autoassociative neural networks. *AIChE J.* **1991**, *37*, 233–243. [CrossRef]

104. Schölkopf, B.; Smola, A.; Müller, K.-R. Nonlinear Component Analysis as a Kernel Eigenvalue Problem. *Neural Comput.* **1998**, *10*, 1299–1319. [CrossRef]
105. Coifman, R.R.; Lafon, S.; Lee, A.B.; Maggioni, M.; Nadler, B.; Warner, F.; Zucker, S.W. Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps. *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 7426–7431. [CrossRef]
106. Coifman, R.R.; Lafon, S. Diffusion maps. *Appl. Comput. Harmon. Anal.* **2006**, *21*, 5–30. [CrossRef]
107. de la Portey, J.; Herbsty, B.M.; Hereman, W.; van der Walty, S.J. An Introduction to Diffusion Maps. In Proceedings of the The 19th Symposium of the Pattern Recognition Association of South Africa (PRASA 2008), Cape Town, South Africa, 27–28 November 2008.
108. Ferguson, A.L.; Zhang, S.; Dikiy, I.; Panagiotopoulos, A.Z.; Debenedetti, P.G.; James Link, A. An experimental and computational investigation of spontaneous lasso formation in microcin J25. *Biophys. J.* **2010**, *99*, 3056–3065. [CrossRef] [PubMed]
109. Kim, S.B.; Dsilva, C.J.; Kevrekidis, I.G.; Debenedetti, P.G. Systematic characterization of protein folding pathways using diffusion maps: Application to Trp-cage miniprotein. *J. Chem. Phys.* **2015**, *142*, 085101. [CrossRef]
110. Trstanova, Z.; Leimkuhler, B.; Lelièvre, T. Local and global perspectives on diffusion maps in the analysis of molecular systems. *Proc. R. Soc. A Math. Phys. Eng. Sci.* **2020**, *476*, 20190036. [CrossRef] [PubMed]
111. Hus, J.C.; Bruschweiler, R. Principal component method for assessing structural heterogeneity across multiple alignment media. *J. Biomol. NMR* **2002**, *24*, 123–132. [CrossRef] [PubMed]
112. Howe, P.W. Principal components analysis of protein structure ensembles calculated using NMR data. *J. Biomol. NMR* **2001**, *20*, 61–70. [CrossRef]
113. Yang, L.W.; Eyal, E.; Bahar, I.; Kitao, A. Principal component analysis of native ensembles of biomolecular structures (PCA_NEST): Insights into functional dynamics. *Bioinformatics* **2009**, *25*, 606–614, Erratum in *Bioinformatics* **2009**, *25*, 2147–2147. [CrossRef]
114. Sakuraba, S.; Kono, H. Spotting the difference in molecular dynamics simulations of biomolecules. *J. Chem. Phys.* **2016**, *145*, 074116. [CrossRef]
115. Wang, H.; Yan, S.; Xu, D.; Tang, X.; Huang, T. Trace Ratio vs. Ratio Trace for Dimensionality Reduction. In Proceedings of the 2007 IEEE Conference on Computer Vision and Pattern Recognition, 17–22 June 2007; pp. 1–8.
116. Ngo, T.T.; Bellalij, M.; Saad, Y. The Trace Ratio Optimization Problem. *SIAM Rev.* **2012**, *54*, 545–569. [CrossRef]
117. Peters, J.H.; de Groot, B.L. Ubiquitin dynamics in complexes reveal molecular recognition mechanisms beyond induced fit and conformational selection. *PLoS Comput. Biol.* **2012**, *8*, e1002704. [CrossRef] [PubMed]
118. Ahmad, M.; Helms, V.; Kalinina, O.V.; Lengauer, T. Relative Principal Components Analysis: Application to Analyzing Biomolecular Conformational Changes. *J. Chem. Theory Comput.* **2019**, *15*, 2166–2178. [CrossRef] [PubMed]
119. Molgedey, L.; Schuster, H.G. Separation of a Mixture of Independent Signals Using Time-Delayed Correlations. *Phys. Rev. Lett.* **1994**, *72*, 3634–3637. [CrossRef] [PubMed]
120. Naritomi, Y.; Fuchigami, S. Slow dynamics in protein fluctuations revealed by time-structure based independent component analysis: The case of domain motions. *J. Chem. Phys.* **2011**, *134*, 065101. [CrossRef]
121. Naritomi, Y.; Fuchigami, S. Slow dynamics of a protein backbone in molecular dynamics simulation revealed by time-structure based independent component analysis. *J. Chem. Phys.* **2013**, *139*, 215102. [CrossRef]
122. Mori, T.; Saito, S. Dynamic heterogeneity in the folding/unfolding transitions of FiP35. *J. Chem. Phys.* **2015**, *142*, 135101. [CrossRef]
123. Takano, H.; Miyashita, S. Relaxation Modes in Random Spin Systems. *J. Phys. Soc. Jpn.* **1995**, *64*, 3688–3698. [CrossRef]
124. Hirao, H.; Koseki, S.; Takano, H. Molecular Dynamics Study of Relaxation Modes of a Single Polymer Chain. *J. Phys. Soc. Jpn.* **1997**, *66*, 3399–3405. [CrossRef]
125. Koseki, S.; Hirao, H.; Takano, H. Monte Carlo Study of Relaxation Modes of a Single Polymer Chain. *J. Phys. Soc. Jpn.* **1997**, *66*, 1631–1637. [CrossRef]
126. Mitsutake, A.; Iijima, H.; Takano, H. Relaxation mode analysis of a peptide system: Comparison with principal component analysis. *J. Chem. Phys.* **2011**, *135*, 164102. [CrossRef] [PubMed]
127. Mitsutake, A.; Takano, H. Relaxation mode analysis and Markov state relaxation mode analysis for chignolin in aqueous solution near a transition temperature. *J. Chem. Phys.* **2015**, *143*, 124111. [CrossRef] [PubMed]
128. Karasawa, N.; Mitsutake, A.; Takano, H. Two-step relaxation mode analysis with multiple evolution times applied to all-atom molecular dynamics protein simulation. *Phys. Rev. E* **2017**, *96*, 062408. [CrossRef] [PubMed]
129. Schultze, S.; Grubmüller, H. Time-Lagged Independent Component Analysis of Random Walks and Protein Dynamics. *J. Chem. Theory Comput.* **2021**, *17*, 5766–5776. [CrossRef] [PubMed]
130. Morishita, T. Time-dependent principal component analysis: A unified approach to high-dimensional data reduction using adiabatic dynamics. *J. Chem. Phys.* **2021**, *155*, 134114. [CrossRef]
131. Perez-Hernandez, G.; Paul, F.; Giorgino, T.; De Fabritiis, G.; Noe, F. Identification of slow molecular order parameters for Markov model construction. *J. Chem. Phys.* **2013**, *139*, 015102. [CrossRef]
132. Scherer, M.K.; Trendelkamp-Schroer, B.; Paul, F.; Perez-Hernandez, G.; Hoffmann, M.; Plattner, N.; Wehmeyer, C.; Prinz, J.H.; Noe, F. PyEMMA 2: A Software Package for Estimation, Validation, and Analysis of Markov Models. *J. Chem. Theory Comput.* **2015**, *11*, 5525–5542. [CrossRef]
133. Harrigan, M.P.; Sultan, M.M.; Hernandez, C.X.; Husic, B.E.; Eastman, P.; Schwantes, C.R.; Beauchamp, K.A.; McGibbon, R.T.; Pande, V.S. MSMBuilder: Statistical Models for Biomolecular Dynamics. *Biophys. J.* **2017**, *112*, 10–15. [CrossRef]

134. Schwantes, C.R.; Pande, V.S. Modeling molecular kinetics with tICA and the kernel trick. *J. Chem. Theory Comput.* **2015**, *11*, 600–608. [CrossRef]
135. Husic, B.E.; Pande, V.S. Markov State Models: From an Art to a Science. *J. Am. Chem. Soc.* **2018**, *140*, 2386–2396. [CrossRef]
136. Wang, X.; Unarta, I.C.; Cheung, P.P.; Huang, X. Elucidating molecular mechanisms of functional conformational changes of proteins via Markov state models. *Curr. Opin. Struct. Biol.* **2021**, *67*, 69–77. [CrossRef]