The Protein Journal (2023) 42:181–191
https://doi.org/10.1007/s10930-023-10113-9

A Retrospective on the Development of Methods for the Analysis of Protein Conformational Ensembles

Steven Hayward[1]

Accepted: 5 April 2023 / Published online: 19 April 2023

Abstract
Analysing protein conformational ensembles whether from molecular dynamics (MD) simulation or other sources for functionally relevant conformational changes can be very challenging. In the nineteen nineties dimensional reduction methods were developed primarily for analysing MD trajectories to determine dominant motions with the aim of understanding their relationship to function. Coarse-graining methods were also developed so the conformational change between two structures could be described in terms of the relative motion of a small number of quasi-rigid regions rather than in terms of a large number of atoms. When these methods are combined, they can characterize the large-scale motions inherent in a conformational ensemble providing insight into possible functional mechanism. The dimensional reduction methods first applied to protein conformational ensembles were referred to as Quasi-Harmonic Analysis, Principal Component Analysis and Essential Dynamics Analysis. A retrospective on the origin of these methods is presented, the relationships between them explained, and more recent developments reviewed.

Keywords Principal Component Analysis · Essential Dynamics · Quasi-Harmonic Analysis · Collective motions · Domain motions

1 Introduction

It is now a long-established fact that protein function is intimately linked to protein conformational change as demonstrated by the solved structures of proteins in multiple functional states. For example, they show how the binding of a ligand can induce a conformational change. If we have only a single structure of a protein, however, there are a number of computational techniques that we can use to model dynamics and so gain insight into functionally related motions. The most popular and accurate is molecular dynamics (MD) simulation which is usually performed with the protein immersed in a bath of water molecules, and there are several well-known MD packages for this purpose, e.g., GROMACS [4], AMBER [5], CHARMM [6] and NAMD [7]. Other methods include Normal Mode Analysis (NMA) [8–14] and Monte Carlo (MC) sampling [15–21]. MD and MC give a trajectory for every atom, and the challenge is to extract functionally relevant motions hidden within the noisy trajectories of a large number of atoms. Experimental techniques such as NMR, X-ray crystallography, and more recently cryo-electron microscopy [22] can also provide us with conformational ensembles.

This review is mainly focussed on what is broadly known as Principal Component Analysis (PCA) in application to protein conformational ensembles. PCA is a general dimensional reduction method that is widely applied in many fields and is used on high-dimensional data with the aim of representing the data as faithfully as possible using fewer dimensions. In application to conformational ensembles, it went under different names depending on context. In Quasi-Harmonic Analysis (QHA), the context is its relationship to NMA which assumes that a protein behaves like a harmonic oscillator. "Essential Dynamics Analysis" (EDA), developed in the Berendsen group by Amadei, Linnsen and Berendsen [1], is also PCA but captures in its terminology an important feature of protein motion, which is that most of the motion occurs in a subspace, called the "essential subspace" spanned by a very small number of dimensions. The emphasis of that paper and its terminology helped convey

✉ Steven Hayward
steven.hayward@uea.ac.uk

[1] Laboratory for Computational Biology, School of Computing Sciences, University of East Anglia, Norwich, UK

this feature of protein dynamics to the MD community better than other publications [23–25] at the time even though they had similar findings.

By measuring the overlap of their essential subspaces, PCA offers a good way to compare results from two simulations or from a simulation and a protein conformational ensemble from experiment. These methods can also be used to establish the stability of the essential subspace from a single simulation, for example by dividing the trajectory into two parts.

Sometimes PCA on a protein trajectory can reveal a single very dominant mode of motion. Even in this case, the dominating mode's motion can be very complex and difficult to understand, specifying in a collective fashion the movement of all the atoms. However, depending on the nature of the motion, coarse-graining methods can be applied that reduce the description from an all atom one to one that of a few quasi-rigid regions or domains. Thus, combining PCA with a coarse-graining method can produce, from a highly complex and noisy trajectory, a depiction that is easy to comprehend.

In addition to reviewing the origins of PCA, EDA and QHA for the analysis of protein ensembles and the connections between them, some more recent variants of PCA that have been applied in this area will be reviewed.

2 First Applications of PCA, QHA and EDA to Protein Conformational Ensembles

In this section we review how PCA has been applied to proteins and try to give a perspective on the variants of PCA that arose in its application to protein conformational ensembles. As there has been a recent excellent review by Kitao [26] on PCA in application to protein dynamics, that review can be referred to for further details.

PCA is a multi-variate technique that can be applied to a wide variety of data and it predates its application to protein dynamics by over 100 years [27]. Its first application to protein dynamics was framed in the context of NMA and was called "Quasi-Harmonic Analysis". In order to appreciate QHA it is necessary to explain NMA.

2.1 QHA and Its Relationship to NMA

NMA is a harmonic method that models protein dynamics at physiological temperatures using the parabolic approximation of the conformational energy surface at a single energy minimum (for details see the early papers on NMA [8, 10, 14] or some later reviews [9, 11–13, 26, 28]). To simplify the kinetic energy term in the Lagrangian, NMA in Cartesian coordinate space uses mass-weighted atomic displacements ($\Delta q_i = \sqrt{m_j}\Delta x_j$, $\Delta q_{i+1} = \sqrt{m_j}\Delta y_j$, $\Delta q_{i+2} = \sqrt{m_j}\Delta z_j$) where m_j is the mass of the jth atom ($j = 1,N$, where N is the total number of atoms) and $\Delta x_j, \Delta y_j, \Delta z_j$ give its displacement from its position at the energy minimum conformation. Performing NMA gives a set of eigenvectors $\boldsymbol{v}_k$ ($3N\times 1$ column vectors, $k = 1,3N\text{-}6$) that define shape changing patterns of atomic displacements; the remaining 6 define the external degrees of freedom (translational and rotational degrees of freedom of the whole molecule) that have eigenvalues equal to zero. The form of these external eigenvectors is such that they satisfy the Eckart conditions [29–31]. The eigenvectors define collective variables (normal mode variables):

$$\sigma_k = \boldsymbol{v}_k^t \Delta \boldsymbol{q} = \sum_{i=1}^{3N} v_{ik}\Delta q_i, \tag{1}$$

where t denotes the transpose, $\Delta\mathbf{q} = (\Delta q_1 \Delta q_2 \ldots \Delta q_{3N})^t$, the $3N\times 1$ column vector of mass-weighted atomic displacements, and v_{ik} is the ith element of the eigenvector, $\boldsymbol{v}_k$. Equation (1) shows the collective nature of the normal mode variables in being a linear sum of the atomic displacements. The normal mode variables behave as independent harmonic oscillators, each with an angular frequency given by its associated eigenvalue, ω_k^2. Statistical mechanics for a harmonic oscillator in thermal equilibrium gives:

$$\left\langle \sigma_k^2 \right\rangle = \frac{k_B T}{\omega_k^2}, \tag{2}$$

where k_B is Boltzmann's constant, and T the absolute temperature. Also, one can show that [30]:

$$\sum_{j=1}^{N} m_j \left(\left\langle \Delta x_j^2 \right\rangle + \left\langle \Delta y_j^2 \right\rangle + \left\langle \Delta z_j^2 \right\rangle \right) = \sum_{k=1}^{3N-6} \left\langle \sigma_k^2 \right\rangle = \sum_{k=1}^{3N-6} \frac{k_B T}{\omega_k^2}. \tag{3}$$

The left-hand side of this equation is a mass-weighted total mean-square displacement and is a measure of the total overall motion of the protein in thermal equilibrium. It shows that the lowest frequency normal modes have the largest contribution whatever the frequency distribution. Frequency distributions on the small globular proteins that NMA was first performed, showed how the contribution of a relatively small number of low-frequency normal modes dominated the total mean square fluctuation of the whole protein. This led to the concept of the "important subspace" [28, 30], which is the subspace defined by the lowest frequency normal modes in which most of the motion occurs. NMA also allows one to calculate the variance–covariance matrix for the mass-weighted atomic coordinate displacements as:

$$\left\langle \Delta q_i \Delta q_j \right\rangle = k_B T \sum_{k=1}^{3N-6} \frac{v_{ik} v_{jk}}{\omega_k^2}, \tag{4}$$

or in matrix form:

$$\boldsymbol{C} = \boldsymbol{V}\boldsymbol{\lambda}\boldsymbol{V}^t, \tag{5}$$

where $\boldsymbol{C}$ is the variance–covariance matrix with elements $\langle \Delta q_i \Delta q_j \rangle$, $\boldsymbol{V}$ is the eigenvector matrix, the kth column of which is $\boldsymbol{v}_k$, and $\boldsymbol{\lambda}$, a diagonal matrix with elements, $\lambda_k = \frac{k_B T}{\omega_k^2}$ which are the mean-square displacement or mean-square fluctuation (msf) of the normal mode variables [see Eq. (2)]. Equation (5) reveals the origin of QHA as it shows how one can derive NMA eigenvectors and eigenvalues from the variance–covariance matrix. This would mean that if an MD simulation were performed for a system with a single parabolic energy well, then provided it were sufficiently long (see below for more on convergence), good approximations to NMA eigenvectors and eigenvalues could be determined.

In NMA time plays a central role as one is solving Newton's equations of motion. However, in performing QHA time is not explicitly involved, and it can therefore be applied to protein conformational ensembles where there is no time ordering of the conformations, e.g., ensembles of crystallographic structures.

Although NMA could predict atomic B-factors well and the lowest frequency normal modes were plausible in that for proteins like lysozyme they produced the expected domain motion [32], the assumption of harmonicity is in a strict sense wrong as it is known that the state point (the point in the 3N-6 dimensional space that represents the conformation) visits multiple energy minima. This was demonstrated in early experimental observations [33] and MD simulations [34] and a large body of work has since supported this. In contrast to NMA, with MD simulations the state point can move from minimum to minimum giving a more realistic simulation of protein dynamics. Despite this it is still possible to do QHA by calculating the variance–covariance matrix irrespective of the nature of the conformational energy surface. To do this one first has to remove the external degrees of freedom which is done by performing mass-weighted least-squares best fits of the conformations to a reference structure, e.g., the starting structure. This mass-weighting is important as it can be shown [31] that in doing so, the Eckart conditions for removal of the external degrees of freedom are satisfied. For NMA, displacements are calculated from the minimum energy structure which would be the same as the average structure for motion on a truly parabolic energy surface. To perform QHA the mass-weighted displacements used to calculate the variance–covariance matrix are calculated from the simulation average structure. It can be performed as follows. After removing the external degrees of freedom and calculating the average structure, a matrix of each atom's mass-weighted displacement from its average position, at each time frame can be constructed:

$$\boldsymbol{Q} = \left(\Delta\boldsymbol{q}_1 \Delta\boldsymbol{q}_2 \ldots \Delta\boldsymbol{q}_l \ldots \Delta\boldsymbol{q}_L\right), \tag{6}$$

where, as above, $\Delta\boldsymbol{q}_l$ is a column vector of the mass-weighted atomic displacements at time frame l, and L is the number of time frames saved from the simulation. The variance–covariance matrix can then be calculated as:

$$\boldsymbol{C} = \frac{1}{L}\boldsymbol{Q}\boldsymbol{Q}^t. \tag{7}$$

Diagonalisation of $\boldsymbol{C}$ gives the eigenvector matrix $\boldsymbol{V}$ and eigenvalue matrix, $\boldsymbol{\lambda}$, as in Eq. (5) above. In NMA one is interested in the lowest frequency motions as they produce the largest fluctuations [see Eq. (2)] and so one would sort the eigenvalues from lowest to highest, whereas in doing QHA one sorts the eigenvalues from highest to lowest given that the eigenvalues directly give the msf's of the quasi-harmonic mode variables. The projection of the trajectory into the space of the first n QHA coordinates is given by:

$$\begin{pmatrix} \boldsymbol{\sigma}_1 \\ \boldsymbol{\sigma}_2 \\ \vdots \\ \boldsymbol{\sigma}_i \\ \vdots \\ \boldsymbol{\sigma}_n \end{pmatrix} = \left(\boldsymbol{v}_1 \boldsymbol{v}_2 \ldots \boldsymbol{v}_i \ldots \boldsymbol{v}_n\right)^t \boldsymbol{Q}, \tag{8}$$

where $\boldsymbol{\sigma}_i$ is a $1 \times L$ row vector giving the atomic displacements at each frame projected onto the ith QHA coordinate.

The earliest application to protein dynamics of QHA was by Karplus and Kushick [35] in order to estimate the configurational entropy. Later applications of QHA by Levy et al. [36, 37] to butane and BPTI concentrated on the frequency distributions derived from the eigenvalues rather than inspecting motions along eigenvectors.

2.2 First Applications of PCA and Essential Dynamics Analysis to MD Trajectories

QHA is in fact PCA on the mass-weighted coordinates but framed as an inverse procedure to NMA that can be applied to MD and MC simulations.

In the early nineteen nineties, papers [1, 23–25] appeared that analysed MD trajectories using PCA that aligned closer to the origins of PCA as a geometrical method for finding the orthogonal transformation that best represents a distribution of points using fewer dimensions. Many of these papers did not directly frame their work in terms of QHA even if they did use mass-weighted displacements, preferring to use the term PCA, and others did not mass-weight the displacements thus breaking the formal connection to QHA.

These papers showed plots of the cumulative msf's with principal coordinate number (ordered from largest eigenvalue to smallest). They showed the dominance of a small

number of the first principal coordinates in their contribution to the total msf; this dominance being more dramatic than seen with NMA. Of particular impact was study carried out in the Berendsen group by Amadei et al. [1], where PCA was framed as an "Essential Dynamics Analysis". The analysis was also performed on C_α atoms only, unique at the time as others used all the atoms. A particular emphasis was put on the small size of the subspace within which protein dynamics is largely confined, and the terms such "essential" and "near constraint" (in the sense of an effective constraint) served to convey this message very well. Figure 1 shows the original plots from Amadei et al. for the relative cumulative fluctuation against eigenvector number. An early application that arose from EDA was a new sampling technique that accelerates sampling within the essential subspace [38]. In the Gō group, the focus was on the dynamical behaviour of collective motions [24, 39]. In particular Langevin mode analysis [40] was performed which assumes the motion can be modeled as a harmonic oscillator in a viscous fluid.

Projection onto the first two principal components is particularly informative as clusters can be seen. For example, a comparison of MD trajectories [24, 25] in vacuo and in explicit water showed how the presence of water

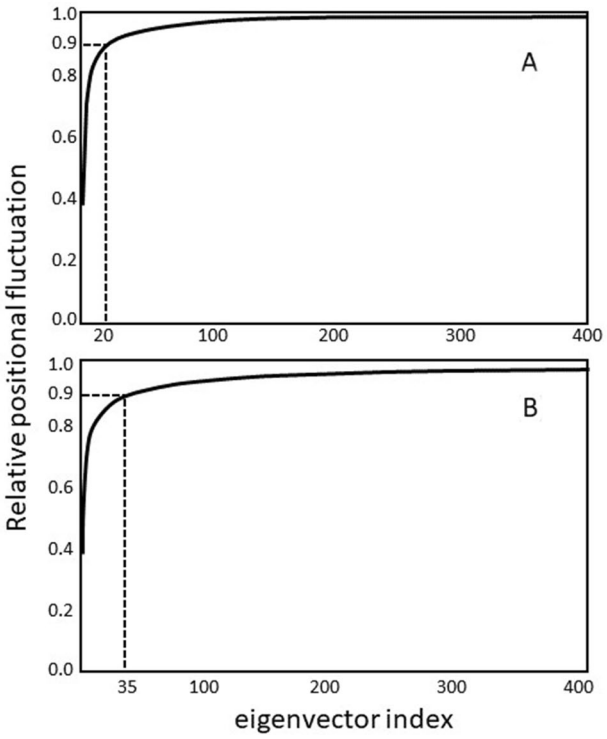

Fig. 1 Reproduced from Fig. 2 in Amadei et al. [1] showing the relative cumulative fluctuation against eigenvector number for an essential dynamics analysis on a 900 ps solvent MD simulation of lysozyme. This demonstrates the dominance of a relatively small number (out of a possible 3792) of "essential" eigenvectors. **A** Cα atoms only. **B** All atoms

created small local clusters in the projected trajectories. Such projections can also show clusters that might relate to functionally related stable states in the conformational landscape. Projection of MD trajectories on to individual principal components with the highest msf's to produce "probability distributions" (histograms of population density) clearly demonstrated the non-harmonic nature of protein dynamics. In this regard the QHA approach is particularly useful as one can compare results directly to those obtain from NMA. One would not expect NMA eigenvectors to be well aligned to QHA eigenvectors of an MD trajectory as NMA is performed in a single energy minimum, whereas in MD the state point visits multiple energy minima. Nevertheless, NMA can be used to distinguish anharmonic and harmonic QHA modes. If we denote the NMA eigenvectors, $\boldsymbol{w}_i$, and the QHA eigenvectors, $\boldsymbol{v}_k$, then we can project the NMA msf onto the kth QHA mode using:

$$\lambda_k^{har} = k_B T \sum_{i=1}^{3N-6} \frac{\left(\boldsymbol{w}_i^t \boldsymbol{v}_k\right)^2}{\omega_i^2}. \tag{9}$$

The "anharmonicity factor" [41] for each QHA mode is defined as:

$$\mu_k = \sqrt{\frac{\lambda_k}{\lambda_k^{har}}}, \tag{10}$$

where λ_k is the msf of the kth QHA mode (its eigenvalue) from the MD simulation. If $\mu_k = 1$ then it means that the msf derived from all the normal modes projected onto the kth QHA mode matches that from the MD trajectory projected onto the kth QHA mode. The QHA modes with $\mu_k = 1$ were referred to as harmonic and their probability distributions showed the expected Gaussian form. Those QHA modes with $\mu_k > 1$ were suggested to be anharmonic as their msf's could not be reproduced by NMA. Those with the largest values of μ_k had multi-modal distributions consistent with a state point moving on an energy surface with multiple minima. Studies revealed that QHA modes separate into low number modes that are anharmonic, and high number modes that are harmonic. For BPTI simulated in vacuo only 12% of the mode were anharmonic, but they contributed 98% to the total msf [41]. A comparable result was found for a simulation of lysozyme in water [42]. An interesting finding was that the larger the msf of a mode, the larger its anharmonicity factor, suggesting that large-scale movements in proteins are those that derive principally from anharmonic, minima jumping events. This analysis led to a variant of QHA, called the Jumping Among Minima or JAM model [42] that can separate the contributions to the variance–covariance matrix into those that arise from fluctuations within minima and those that arise from fluctuations between minima. Using

this model, it was possible to gain insight into the structure of the energy surface for a protein, revealing it to have a hierarchical nature. Amadei et al. [43] developed a double diffusion model for the kinetics of "essential" coordinates that combined motions within energy minima with jumps between energy minima, the former having a higher diffusion constant than that of the latter.

PCA, being a statistical method is subject to sampling errors. Hess [44] a member of the Berendsen group at the time, showed that applying PCA to random diffusion in a high-dimensional space can give the impression of underlying correlations even when there are none. It was shown that trajectories from random diffusion projected onto the dominant PCA mode variables have a cosine form. This can be a strong indication of non-convergence, but its presence alone does not necessarily indicate an unstable subspace. If the results of PCA on protein MD trajectories are to have meaning, then the subspaces of the dominant PCA modes should not vary dramatically between two different portions of a trajectory from a single simulation of a protein in thermal equilibrium, or indeed between two different equilibrium simulations of the same protein in the same state. Thus, one can quantify the stability of the subspace by measuring the overlap of the two subspaces. Labelling the two trajectories, or two portions of a single trajectory, a and b, the root mean-square inner product (rmsip), is a measure which directly quantifies the overlap, $O_M^{a,b}$, of two subspaces:

$$O_M^{a,b} = \sqrt{\frac{1}{M}\sum_{i=1}^{M}\sum_{j=1}^{M}\left(\boldsymbol{v}_{a,i}^t \boldsymbol{v}_{b,j}\right)^2}, \tag{11}$$

where $\boldsymbol{v}_{a,i}$ and $\boldsymbol{v}_{b,j}$ are the ith and jth eigenvectors from the PCA analyses of trajectories, a and b, and M is the dimension of the subspaces. $O_M^{a,b}$ is equal to 1.0 for fully overlapping subspaces. An early study [45] using state-of-the-art MD simulations at the time on G-actin (470 ps), suggested that PCA did not give a stable dominant subspace. This was established by performing PCA (375 C_α atoms) on each of the two 235 ps halves of the trajectory and comparing the subspaces using a measure related to the rmsip. However, a slightly later study [46] using 2 ns MD simulations on protein L and Cytochrome c551 found the essential subspaces to be stable. A much more recent study applying PCA to trajectories from multiple MD simulations on BPTI and lysozyme of ten's of nanoseconds duration, has served to confirm the stability of subspaces defined by dominant PCA modes [47].

3 Recent Developments based on PCA

Since the emergence of PCA as a powerful method for analysing protein trajectories, many other variants and applications have since been developed.

3.1 Linear Response

Linear response can be used to determine the conformational response of a system under external forces. In application to protein–ligand binding it can be stated as follows: the equilibrium fluctuations of the protein in the absence of the ligand, can be used to approximate the response of the protein due to forces of interaction with the ligand. It was shown by Ikeguchi et al. [48] to reproduce quite accurately known ligand-induced conformational changes for a selection of proteins. The basic formula is:

$$\Delta\boldsymbol{r} = \frac{1}{k_B T}\boldsymbol{C}\boldsymbol{f}, \tag{12}$$

where $\boldsymbol{f}$ is the $3N\times 1$ force vector, giving the force on each protein atom from the ligand, $\Delta\boldsymbol{r}$ is the $3N\times 1$ displacement vector, giving the displacement of each protein atom, and $\boldsymbol{C}$ is the $3N\times 3N$ variance–covariance matrix (here not mass-weighted). This approach has been used in the interactive docking tool, DockIT [3, 49, 50], for docking a ligand to a protein receptor. DockIT enables the user to control the ligand position and orientation using either a keyboard and mouse, a haptic device, or in VR using hand-held controllers. To model interaction forces it uses GROMACS [4] topology files generated using the *pdb2gmx* command and GROMACS itp files containing the non-bonded interaction parameters for GROMOS and AMBER force fields. Interactive docking is an interesting application as calculations have to be evaluated within real time limits (<30 ms for graphics, <2 ms for haptics) in order to produce a smooth realistic experience. Evaluation of $\Delta\boldsymbol{r}$ in Eq. (12) requires $9N^2$ multiplications which even using a modern graphics card cannot be achieved in real-time when using a haptic device and/or with a large protein. Diagonalisation of $\boldsymbol{C}$ enables the following approximation to Eq. (12):

$$\Delta\boldsymbol{r} \approx \frac{1}{k_B T}\boldsymbol{V}_M \boldsymbol{\lambda}_M \boldsymbol{V}_M^t \boldsymbol{f}, \tag{13}$$

where $\boldsymbol{V}_M$ is the $3N\times M$ eigenvector matrix containing the first M dominant PC modes, and $\boldsymbol{\lambda}_M$ the $M\times M$ diagonal eigenvalues matrix of corresponding eigenvalues. Using Eq. (13) to get an approximation to $\Delta\boldsymbol{r}$ requires $M(6N+1)$ multiplications (multiplying from right to left). The idea behind this is that even though M might have to be small in order to satisfy time and memory constraints, the approximation in Eq. (13) may still be very good as most of the

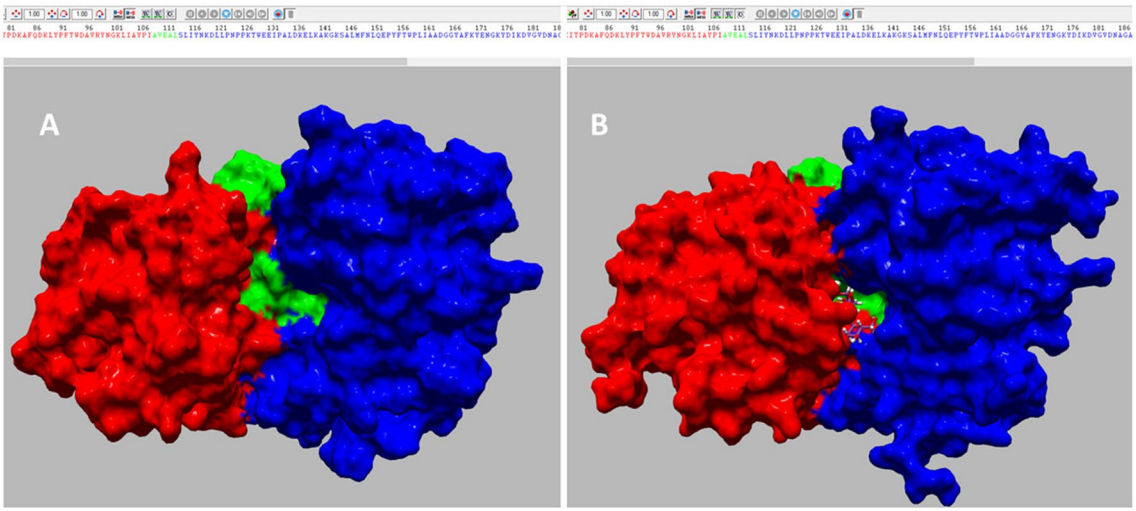

Fig. 2 Domain movement in MBP from docking maltose to MBP with DockIT[3] using the linear response model [see Eq. (13)] to model the conformational change upon maltose binding. Only 26 eigenvalues and eigenvectors were used which were derived from a 100 ns explicit solvent MD simulation of MBP in its maltose free state. Colouring shows domains (red and blue) and hinge bending regions (green) as assigned by DynDom for the movement between the maltose-free (PDB: 1OMP) and maltose-bound structure (PDB: 1ANF). **A** The relaxed structure of MBP without maltose. **B** A closed domain MBP structure with maltose (ball-and-stick) docked into its binding site

fluctuation occurs within the important or essential subspace. Indeed, in the case of maltose binding to maltose binding protein (MBP), where $\boldsymbol{C}$ was calculated from a 100 ns explicit solvent MD simulation of MBP, only about 3% of the total number of eigenvectors could be used, but these represented nearly 90% of the total fluctuation [50]. Figure 2 shows the result of docking maltose to MBP using DockIT when only 26 of the 17,205 eigenvectors are used, i.e., just 0.15%. Despite the very small size of the subspace, docking maltose into the interdomain cleft resulted in a domain movement that matched very well the domain movement between the crystallographic unbound and maltose-bound structures [3].

3.2 PCA in Dihedral Angle Space

PCA is not limited to Cartesian coordinates, although it is the most straightforward to perform. The exact tertiary structure of a protein will be specified by all its so-called internal variables which would be the whole set of bond lengths, bond angles and dihedral angles. Bond lengths and angles are rather constrained in comparison to dihedral angles and so it is common to consider only the rotatable dihedrals which reduces the number of variables compared to the number of Cartesian coordinates by about a factor of eight for proteins. Using internal variables also means no fitting to a fixed structure is necessary in performing PCA. NMA can also be carried out using only the dihedral angles and in doing so one needs to be able to convert dihedral angle changes to Cartesian coordinate displacements via a Jacobian that is derived to satisfy the Eckart conditions and calculated for the energy minimum structure. Omori et al. [51] showed how to perform dihedral angle PCA in an analogous procedure but the Jacobian matrix, $\boldsymbol{L}$, for the transformation between dihedral angle changes and Cartesian coordinate displacements is calculated at the average MD structure rather than at the energy minimum structure. In the linear approximation the relationship between atomic displacements and angle changes is given by:

$$\Delta \boldsymbol{r} = \boldsymbol{L}\Delta\boldsymbol{\theta}, \tag{14}$$

where $\Delta\boldsymbol{r}$ is the $3N\times 1$ the vector of atomic displacements, $\Delta\boldsymbol{\theta}$ is the $M\times 1$ vector of dihedral angle changes, and $\boldsymbol{L}$ is the $3N\times M$ Jacobian matrix. Equation (14) be can used to define the dihedral angle variance–covariance matrix as:

$$\boldsymbol{C}_\theta = \left(\boldsymbol{L}^t\boldsymbol{C}^{-1}\boldsymbol{L}\right)^{-1} \tag{15}$$

Comparison of $\boldsymbol{C}_\theta$ to the variance covariance matrix performed directly on dihedral angles changes (from their average) revealed motions corresponding to compensating dihedral angles changes that maintain the overall structure of a protein and which were referred to as "latent dynamics" [51, 52]. In a different context these motions have been called path-preserving motions, [53] specific examples being the backrub motion [54] and the peptide plane flip [55]. This approach has also been used with

some success to model the linear response of proteins upon binding a ligand [51].

There are other approaches that use dihedral angles such as that by Stocks and co-workers [56, 57] who build a variance–covariance matrix constructed from both the sine and cosine of each dihedral angle, so the variance–covariance matrix is order 2 M. This "dPCA" approach is taken to ensure a proper metric is established in that the distance between two angles is now the distance between their corresponding points on the unit circle. On penta-alanine, where Cartesian coordinate PCA shows a single energy minimum on the first two eigenvectors at the α-helical conformation, for dPCA using the ϕ, ψ dihedrals, multiple minima are seen on the first two eigenvectors [57].

For a review of other dihedral angle-based approaches see the recent review article by Kitao [26].

3.3 Kernel PCA

Kernel PCA is a way to project point distributions onto non-linear coordinates. Kernel methods are often used in machine learning to separate clusters that cannot be separated linearly [58]. It relies on a so-called kernel function that gives the inner product between feature vectors $\Phi(\Delta\mathbf{q})$ which are non-linear functions of the original coordinates. Instead of specifying these functions directly, kernel methods specify them implicitly using the kernel function. The most popular kernel function to use is the Gaussian kernel:

$$k(\mathbf{x},\mathbf{y}) = \Phi(\mathbf{x})^t\Phi(\mathbf{y}) = \exp\left(-\frac{||\mathbf{x}-\mathbf{y}||^2}{2\sigma^2}\right), \tag{16}$$

where σ^2 is a parameter. PCA performed directly on feature vectors would require the solution of the following eigenvector equation:

$$\frac{1}{L}\sum_{l=1}^{L}\Phi\left(\Delta\mathbf{q}_l\right)\Phi\left(\Delta\mathbf{q}_l\right)^t\boldsymbol{v}_j = \lambda_j\boldsymbol{v}_j, \tag{17}$$

where the summation gives the variance–covariance matrix in the feature vector space. It can be shown that $\Phi\left(\Delta\mathbf{q}_l\right)^t\boldsymbol{v}_j$, the projection of the l th feature vector onto the jth kernel PCA eigenvector, can be calculated from the eigenvalues and eigenvectors of the kernel matrix $\boldsymbol{K}_{ij} = k\left(\Delta\mathbf{q}_i, \Delta\mathbf{q}_j\right)$. Thus, one can project the trajectory data onto selected kernel PCA eigenvectors by diagonalizing $\boldsymbol{K}_{ij}$.

Jacob and David [59] have applied kernel PCA to protein trajectories and have provided useful implementation recipes, pointing out common pitfalls. It seems its main use would be in characterizing non-linear motions in protein dynamics and for clustering where clusters are not linearly separable in projections using standard PCA.

Another modern technique is time-lagged independent component analysis (TICA) which is a method to find modes of motions that maximize time-lagged autocorrelation functions derived from a time-lagged covariance matrix. This method has been applied to protein dynamics [60] and can be used for Markov model construction. A recent paper on TICA has discussed convergence of TICA modes from protein trajectories by comparing them to modes derived from random walk trajectories [61].

Other dimensional reduction methods exist although many of them are not widely applied to protein dynamics possibly because they do not appear to dramatically improve upon the results from Cartesian coordinate PCA [62].

4 Comparing Results from PCA on Simulation Trajectories with Experimentally Derived Movements

It is common to compare MD trajectories to experimentally derived movements. It is often the case in X-ray crystallography that the structure of a protein is solved in both a ligand-free state and a ligand-bound state giving the opportunity to compare the results with an MD simulation trajectory. If one performs MD simulation on the ligand-free protein and the trajectory is one of the protein in thermal equilibrium, the effect of the ligand is ignored, and one might wonder about the implication of this. However, the theory of pre-existing populations or conformational selection [63] where the ligand is thought to stabilise a conformation of the ligand-free protein, suggests that this is a good approach. This view is supported by the excellent results from linear response (see above) and impressive results on ubiquitin [64]. It also makes logical sense at least for hinge-bending proteins as their domain motions are clearly encoded in their structures and would therefore be expected to occur in equilibrium, although not necessarily to the extent when binding a ligand.

The dominant PC modes, i.e., those with the highest msf, are also those that are the most "collective", meaning that they involve the displacements of atoms across the whole protein, as opposed to being localized, a feature of the modes with low msf's. It is the high msf collective motions that are the most likely to relate to function [65]. For this reason, it is usual to compare these to experimental derived functional movements.

If one is to compare the results from a PCA of a simulation trajectory to an experimentally determined movement then there are various measures. Consider the case where one has performed a simulation on the ligand-free protein and performed PCA. If there is also a ligand-bound structure, then one measure analogous to the rmsip might be:

$$O_M^{sim,exp} = \sqrt{\sum_{i=1}^{M} \left(\boldsymbol{v}_{sim,i}^t \frac{\boldsymbol{\Delta r}_{exp}}{\left| \boldsymbol{\Delta r}_{exp} \right|} \right)^2}, \tag{18}$$

where $\boldsymbol{\Delta r}_{exp}$ is the $3N \times 1$ vector determined from the movement between the ligand-free and ligand-bound structures, and $\boldsymbol{v}_{sim,i}$ is the ith eigenvector from a PCA of the simulation trajectory. In evaluating $O_M^{sim,exp}$ care should be taken to ensure that the external frame of reference for the calculation of $\boldsymbol{\Delta r}_{exp}$ is the same as for the PCA. $O_M^{sim,exp}$ measures whether the experimentally determined movement lies within the M-dimensional PCA subspace. The maximum value for $O_M^{sim,exp}$ is 1.0 but only values close to 1.0 for small M would indicate a good result, i.e., $O_M^{sim,exp} \to 1.0, M \to 3N - 6$. For some proteins, many structures are available and if there are a sufficient number, then PCA can also be performed and compared to PCA on a simulation trajectory. Again, a rmisp measure can be used to compare results:

$$O_{M,N}^{sim,exp} = \sqrt{\frac{1}{N} \sum_{i=1}^{M} \sum_{j=1}^{N<M} \left(\boldsymbol{v}_{sim,i}^t \boldsymbol{v}_{exp,j} \right)^2}, \tag{19}$$

which measures the overlap between the N-dimensional subspace from the PCA of the experimental structures and the M-dimensional subspace from the PCA of the simulation trajectory. This was done in the Berendsen group on calmodulin [66]and T4 lysozyme [2]. Figure 3 shows the result of projecting 38 crystallographic structures of T4 lysozyme and the trajectories of three separate MD trajectories onto the first two modes from a PCA of the crystallographic structures. The excellent overlap between the subspaces is indicated by $O_{5,1}^{sim,exp} = 0.96$ determined using the first mode from the PCA of the crystallographic structures and the first five PCA modes from a PCA of the combined MD trajectories.

Of course, since the development of these methods, advances mean that longer simulations can be performed, and free energies calculated using sampling techniques.

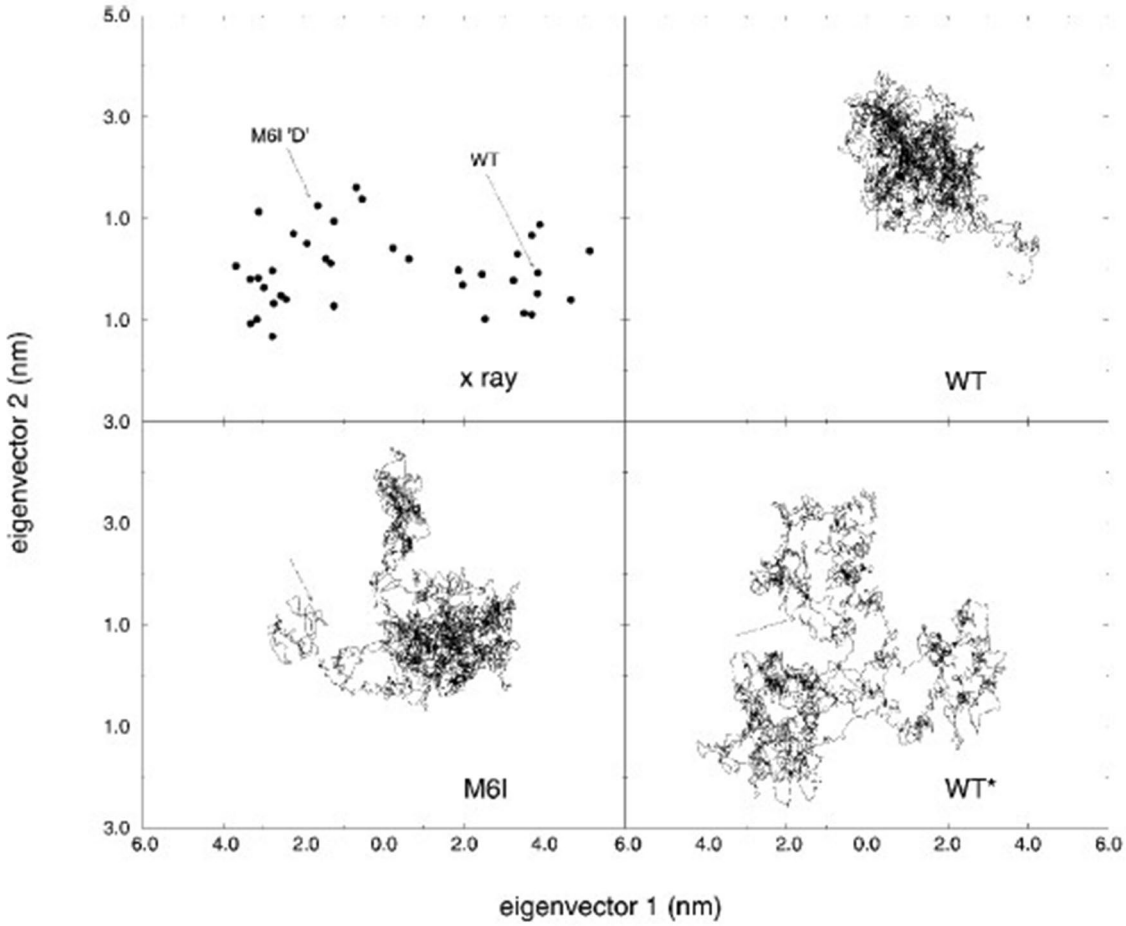

Fig. 3 From de Groot et al. [2] showing the projections onto the 2D plane defined by the first two modes of a PCA of 38 crystallographic structures of T4 lysozyme. The top left plot shows the crystallographic structures themselves and the other plots show the projected trajectories from three independent MD simulations

Free-energy profiles along principal coordinates can be calculated with and without a ligand present using techniques such as umbrella sampling offering a deeper insight into functional movement. Using umbrella sampling along the first PCA coordinate on MBP, it was shown that the apo protein can reach a "semi-closed" metastable conformation near to, but not coincident with the fully-closed maltose-bound structure [67].

5 Visualising and Characterising Motions along Dominant Modes

Even though the movement along a single principal coordinate is linear it usually takes place in a very high-dimensional space representing a large set of atomic movements that results in a conformational change that can be difficult to characterise. A first step to gain insight might be to simply view movements along individual principal coordinates using molecular graphics. This requires the generation of individual structures along the mode. One can do this as follows. First one can project the trajectory onto the selected mode i to find its extent as:

$$\begin{aligned} \sigma_i^{min} = \min_l\left(\boldsymbol{v}_i^t\boldsymbol{Q}\right); \sigma_i^{max} = \max_l\left(\boldsymbol{v}_i^t\boldsymbol{Q}\right) \\ \boldsymbol{r}_i^{min} = \langle \boldsymbol{r} \rangle + \sigma_i^{min}\boldsymbol{v}_i \\ \boldsymbol{r}_i^{max} = \langle \boldsymbol{r} \rangle + \sigma_i^{max}\boldsymbol{v}_i \end{aligned} \tag{20}$$

where $\langle \boldsymbol{r} \rangle$ is the $3N \times 1$ vector of the Cartesian coordinates of the average structure used for PCA and $\boldsymbol{Q}$ is the $3N \times L$ trajectory matrix [see Eq. (6)], here from a non-mass weighted PCA, and $\boldsymbol{r}_i^{min}$ and $\boldsymbol{r}_i^{max}$ are $3N \times 1$ vectors of the Cartesian coordinates of the minimum and maximum extent of the ith mode when projecting the trajectory onto it. To view with molecular graphics, one can generate intermediate structures between $\boldsymbol{r}_i^{min}$ and $\boldsymbol{r}_i^{max}$ so that a smooth motion is seen. It is common to view just the first PC motion ($i = 1$) especially if this is dominant.

Motions in proteins can generally be described in terms of the structural elements that move. It has been found that domain motions form a large class and there are two reasons for this. The first is that most proteins are multi-domain and the second is that by their very nature, links between domains are comparatively weak enabling their relative movement. In the nineteen nineties the number of proteins solved in different conformations corresponding to different functional states increased considerably and tools were developed to determine domains from protein conformational change [68–70]. These methods are coarse-graining methods that group atoms into quasi-rigid regions, often referred to as "dynamic domains". These tools also determine hinge axes that give the rigid-body rotation of one domain relative to the other. All these tools require two conformations. Even though a single PCA mode is linear in the $3N$-6 space, in the 3D space the atomic displacements can be tangential to the circular path taken by an atom in a rotating rigid body. Thus $\boldsymbol{r}_i^{min}$ and $\boldsymbol{r}_i^{max}$ can be used for input to these tools as was done with DynDom for the first two eigenvectors from the PCA on the 38 crystallographic structures of T4 lysozyme [2] revealing a closing motion for the first eigenvector and a twisting motion for the second eigenvector.

6 Conclusions

A retrospective on the development of dimensional reduction methods for the application to protein conformational ensembles has been presented. PCA, a multivariate method that has general application, when first applied to protein ensembles arising from simulation of protein dynamics, had its origin in NMA and was called QHA. In that context its relationship to PCA might not have been appreciated by the community at the time. Later applications framed their approach in terms of the dominance of a small number of modes of motion and focused more on the character of these modes of motion. All but one of these papers referred to the method as PCA or EDA, the latter capturing, in its terminology, the main feature of applying PCA to protein conformational ensembles: the overwhelming dominance of a relatively small number of modes. These studies also demonstrated the anharmonic nature of the dominant modes. Application of PCA can also be used to test the convergence of a simulation, to compare simulation trajectories, or to compare a simulation trajectory to an experimental ensemble.

There are several variants of PCA including dihedral angle space methods and kernel PCA which presents a general non-linear approach. Much is still to be learned about the possible advantages these variants offer over Cartesian coordinate PCA or how they might complement Cartesian coordinate PCA. Once PCA has been performed and the first few modes have been found to be very dominant, then there is still a lot one can do to understand the nature of the movement in an individual mode. Tools exist that can be applied to individual PC modes to help further characterise the nature of the implied movement by describing them in terms of the relative movement of regions or domains rather than individual atoms providing greater insight. In carrying out the analyses reviewed in this paper, the aim should always be to understand function, which almost always involves conformational change and is often synonymous with it.

Author Contributions SH: wrote the manuscript.

Declarations

Competing interests The authors declare no competing interests.

References

1. Amadei A, Linssen ABM, Berendsen HJC (1993) Essential dynamics of proteins. Proteins 17:412–425
2. de Groot BL et al (1998) Domain motions in bacteriophage T4 lysozyme: a comparison between molecular dynamics and crystallographic data. Proteins 31:116–127
3. Iakovou G, Laycock SD, Hayward S (2022) Interactive flexible-receptor molecular docking in virtual reality using DockIT. J Chem Inf Model 62(23):5855–5861. https://doi.org/10.1021/acs.jcim.2c01274
4. Van der Spoel D et al (2005) GROMACS: fast, flexible, and free. J Comput Chem 26(16):1701–1718. https://doi.org/10.1002/jcc.20291
5. Case DA et al (2005) The Amber biomolecular simulation programs. J Comput Chem 26(16):1668–1688. https://doi.org/10.1002/jcc.20290
6. Brooks BR et al (2009) CHARMM: the biomolecular simulation program. J Comput Chem 30(10):1545–1614. https://doi.org/10.1002/jcc.21287
7. Kalé L et al (1999) NAMD2: greater scalability for parallel molecular dynamics. J Comput Phys 151(1):283–312. https://doi.org/10.1006/jcph.1999.6201
8. Brooks B, Karplus M (1983) Harmonic dynamics of proteins: normal modes and fluctuations in bovine pancreatic trypsin inhibitor. Proc Natl Acad Sci USA 80:6571–6575
9. Case DA (1994) Normal mode analysis of protein dynamics. Curr Opin Struct Biol 4:285–290
10. Go N, Noguti T, Nishikawa T (1983) Dynamics of a small globular protein in terms of low-frequency vibrational modes. Proc Natl Acad Sci USA 80:3696–3700
11. Hayward S (2001) Normal mode analysis of biological molecules. In: Becker OM et al (eds) Computational biochemistry and biophysics. Marcel Dekker Inc, New York, pp 153–168
12. Hayward S, de Groot BL (2008) Normal modes and essential dynamics. In: Kukol A (ed) Molecular modelling of proteins. Humana Press, Totowa
13. Kitao A, Go N (1999) Investigating protein dynamics in collective coordinate space. Curr Opin Struct Biol 9(2):164–169
14. Levitt M, Sander C, Stern PS (1983) The normal modes of a protein: Native bovine pancreatic trypsin inhibitor. Int J Quant Chem 10:181–199
15. Hayward S, Kitao A (2015) Monte carlo sampling with linear inverse kinematics for simulation of protein flexible regions. J Chem Theory Comput 11(8):3895–3905. https://doi.org/10.1021/acs.jctc.5b00215
16. Horiuchi T, Go N (1991) Projection of monte carlo and molecular dynamics trajectories onto the normal mode axes: human lysozyme. Proteins 10:106–116
17. Jorgensen WL, Tirado-Rives J (1996) Monte Carlo vs molecular dynamics for conformational sampling. J Phys Chem 100(34):14508–14513. https://doi.org/10.1021/jp960880x
18. Kidera A (1999) Smart Monte Carlo simulation of a globular protein. Int J Quantum Chem 75(3):207–214. https://doi.org/10.1002/(SICI)1097-461X(1999)75:3%3c207::AID-QUA10%3e3.0.CO;2-M
19. Noguti T, Go N (1985) Efficient Monte Carlo method for simulation of fluctuating conformations of native proteins. Biopolymers 24:527–546
20. Uhlherr A (2000) Monte Carlo conformational sampling of the internal degrees of freedom of chain molecules. Macromolecules 33(4):1351–1360. https://doi.org/10.1021/ma9908595
21. Wu MG, Deem MW (1999) Efficient Monte Carlo methods for cyclic peptides. Mol Phys 97(4):559–580
22. Bonomi M, Pellarin R, Vendruscolo M (2018) Simultaneous determination of protein structure and dynamics using cryo-electron microscopy. Biophys J 114(7):1604–1613. https://doi.org/10.1016/j.bpj.2018.02.028
23. Garcia AE (1992) Large-amplitude nonlinear motions in proteins. Phys Rev Lett 68(17):2696–2699
24. Hayward S et al (1993) Effect of solvent on collective motions in globular protein. J Mol Biol 234:1207–1217
25. Kitao A, Hirata F, Go N (1991) The effects of solvent on the conformation and the collective motions of protein: normal mode analysis and molecular dynamics simulation of melittin in water and in vacuum. Chem Phys 158:447–472
26 Kitao A (2022) Principal component analysis and related methods for investigating the dynamics of biological macromolecules. J 5(2):298–317
27 Pearson K (1901) LIII. On lines and planes of closest fit to systems of points in space. Lond Edinb Dublin Philos Mag J Sci 2(11):559–572. https://doi.org/10.1080/14786440109462720
28. Hayward S, Go N (1995) Collective variable description of native protein dynamics. Annu Rev Phys Chem 46:223–250
29. Eckart C (1935) Some studies concerning rotating axes and polyatomic molecules. Phys Rev 47(7):552–558
30. Go N (1990) A theorem on amplitudes of thermal atomic fluctuations in large molecules assuming specific conformations calculated by normal mode analysis. Biophys Chem 35:105–112
31. Noguti T, Go N (1983) A method of rapid calculation of a second derivative matrix of conformational energy for large molecules. J Phys Soc Jpn 52(10):3685–3690
32. Gibrat J, Go N (1990) Normal mode analysis of human lysozyme: Study of the relative motion of the two domains and characterization of the harmonic motion. Proteins 8:258–279
33. Austin RH et al (1975) Dynamics of ligand binding to myoglobin. Biochemistry 14(24):5355–5373
34. Elber R, Karplus M (1987) Multiple conformational states of proteins: a molecular dynamics analysis of myoglobin. Science 235:318–321
35. Karplus M, Kushick JN (1981) Method for estimating the configurational entropy of macromolecules. Macromolecules 14:325–332
36. Levy RM, de la Luz Rojas Olivia, Feisner RA (1984) Quasi-harmonic method for calculating vibrational spectra from classical simulations on multidimensional anharmonic potential surfaces. Journal of Physical Chemistry 88:4233–4238
37. Levy RM et al (1984) Quasi-harmonic method for studying very low frequency modes in proteins. Biopolymers 23:1099–1112
38. Amadei A et al (1996) An efficient method for sampling the essential subspace of proteins. J Biomol Struct Dyn 13:615–625

39. Kitao A, Hirata F, Go N (1993) Effects of solvent on the conformation and the collective motions of a protein. 3. Free energy analysis by the extended RISM theory. J Phys Chem 97:10231–10235
40. Lamm G, Szabo A (1986) Langevin modes of macromolecules. J Chem Phys 85(12):7334–7348
41. Hayward S, Kitao A, Go N (1995) Harmonicity and anharmonicity in protein dynamics: A normal mode analysis and principal component analysis. Proteins 23:177–186
42. Kitao A, Hayward S, Go N (1998) Energy landscape of a native protein: Jumping-among-minima model. Proteins 33:496–517
43. Amadei A et al (1999) A kinetic model for the internal motions of proteins: Diffusion between multiple harmonic wells. Proteins: Struct Funct Bioinform 35(3):283–292
44 Hess B (2002) Convergence of sampling in protein simulations. Phys Rev E. https://doi.org/10.1103/PhysRevE.65.031910
45. Balsera MA et al (1996) Principal component analysis and long time protein dynamics. J Phys Chem 100(7):2567–2572
46. Amadei A, Ceruso MA, Di Nola A (1999) On the convergence of the conformational coordinates basis set obtained by the essential dynamics analysis of proteins' molecular dynamics simulations. Proteins Struct Funct Bioinform 36(4):419–424
47. Cossio-Pérez R, Palma J, Pierdominici-Sottile G (2017) Consistent principal component modes from molecular dynamics simulations of proteins. J Chem Inf Model 57(4):826–834. https://doi.org/10.1021/acs.jcim.6b00646
48 Ikeguchi M et al (2005) Protein structural change upon ligand binding: Linear response theory. Phys Rev Lett. https://doi.org/10.1103/PhysRevLett.94.078102)
49. Iakovou G et al (2020) DockIT: a tool for interactive molecular docking and molecular complex construction. Bioinformatics 36(24):5698–5700. https://doi.org/10.1093/bioinformatics/btaa1059
50. Matthews N et al (2019) Haptic-assisted interactive molecular docking incorporating receptor flexibility. J Chem Inf Model 59(6):2900–2912. https://doi.org/10.1021/acs.jcim.9b00112
51. Omori S et al (2009) Linear response theory in dihedral angle space for protein structural change upon ligand binding. J Comput Chem 30(16):2602–2608. https://doi.org/10.1002/jcc.21269
52. Omori S et al (2010) Latent dynamics of a protein molecule observed in dihedral angle space. J Chem Phys 132(11):115103. https://doi.org/10.1063/1.3360144
53. Nishima W et al (2009) DTA: dihedral transition analysis for characterization of the effects of large main-chain dihedral changes in proteins. Bioinformatics 25(5):628–635. https://doi.org/10.1093/bioinformatics/btp032
54. Davis IW et al (2006) The backrub motion: How protein backbone shrugs when a sidechain dances. Structure 14(2):265–274. https://doi.org/10.1016/j.str.2005.10.007
55. Hayward S (2001) Peptide-plane flipping. Protein Sci 10:2219–2227
56. Altis A et al (2007) Dihedral angle principal component analysis of molecular dynamics simulations. J Chem Phys 126(24):244111. https://doi.org/10.1063/1.2746330
57. Mu Y, Nguyen PH, Stock G (2005) Energy landscape of a small peptide revealed by dihedral angle principal component analysis. Proteins Struct Funct Bioinform 58(1):45–52. https://doi.org/10.1002/prot.20310
58 Shawe-Taylor J, Cristianini N (2004) Kernel methods for pattern analysis. Cambridge University Press, Cambridge
59. David CC, Jacobs DJ (2014) Principal component analysis: a method for determining the essential dynamics of proteins. Methods Mol Biol 1084:193–226. https://doi.org/10.1007/978-1-62703-658-0_11
60. Naritomi Y, Fuchigami S (2013) Slow dynamics of a protein backbone in molecular dynamics simulation revealed by time-structure based independent component analysis. J Chem Phys 139(21):215102. https://doi.org/10.1063/1.4834695
61. Schultze S, Grubmüller H (2021) Time-lagged independent component analysis of random walks and protein dynamics. J Chem Theory Comput 17(9):5766–5776. https://doi.org/10.1021/acs.jctc.1c00273
62. Tribello GA, Gasparotto P (2019) Using dimensionality reduction to analyze protein trajectories. Front Mol Biosci 6:46. https://doi.org/10.3389/fmolb.2019.00046
63. Ma BY et al (2002) Multiple diverse ligands binding at a single protein site: a matter of pre-existing populations. Protein Sci 11(2):184–197. https://doi.org/10.1110/ps.21302lISSN0961-8368
64. Lange OF et al (2008) Recognition dynamics up to microseconds revealed from an RDC-derived ubiquitin ensemble in solution. Science 320(5882):1471–1475. https://doi.org/10.1126/science.1157092
65. Berendsen HJC, Hayward S (2000) Collective protein dynamics in relation to function. Curr Opin Struct Biol 10(2):165–169
66. Spoel DVD et al (1996) Bending of the calmodulin central helix: a theoretical study. Protein Sci 5(10):2044–2053. https://doi.org/10.1002/pro.5560051011
67. Kondo HX et al (2011) Free-energy landscapes of protein domain movements upon ligand binding. J Phys Chem B 115(23):7629–7636. https://doi.org/10.1021/jp111902t
68. Hayward S, Berendsen HJC (1998) Systematic analysis of domain motions in proteins from conformational change: new results on citrate synthase and T4 lysozyme. Proteins 30:144–154
69. Hinsen K, Thomas A, Field MJ (1999) Analysis of domain motions in large proteins. Proteins 34:369–382
70. Wriggers W, Schulten K (1997) Protein domain movements: detection of rigid domains and visualization of hinges in comparisons of atomic coordinates. Proteins 29:1–14