CHAPTER 13

Principal Components Analysis: A Review of its Application on Molecular Dynamics Data

Sarah A. Mueller Stein,[1] Anne E. Loccisano,[1] Steven M. Firestine[2] and Jeffrey D. Evanseck[1]

[1]*Contribution from the Center for Computational Sciences and the Department of Chemistry and Biochemistry Duquesne University, 600 Forbes Avenue, Pittsburgh, PA 15282, USA*

[2]*Eugene Applebaum College of Pharmacy and Health Science, 3134 Eugene Applebaum Building, Wayne State University, 259 Mack Avenue, Detroit, MI 48201, USA*

Contents

1. INTRODUCTION

Molecular dynamics is a proven and powerful tool in the exploration and study of the structure and dynamics that define biomolecular energy

ANNUAL REPORTS IN COMPUTATIONAL CHEMISTRY, VOLUME 2
ISSN: 1574 1400 DOI 10.1016/S1574 1400(06)02013 5

landscapes [1–7]. Technological advances in simulation methodology [8–14] and computer architecture [15,16] have significantly extended both the time scale and length (size) scale of molecular dynamics trajectories [17–22]. Reports in the literature show that the time scale of contemporary molecular dynamics trajectories, carried out with modest computer resources, have increased by roughly four orders of magnitude since the inception of biomolecular simulation. A microsecond simulation has been reported in 2000 [23]; however, after 5 years, this still remains the exception rather than the rule in computational studies. The length scale of practical molecular dynamics simulations has not witnessed such a dramatic increase, since the urgency for larger systems is not as great as for longer time. System sizes have increased by nearly 25 times, where simulations of 25,000 particles are not uncommon.

To put the growth of molecular dynamics simulations into perspective, a rough analogy to Moore's Law [24] can be created. Since the first reported biomolecular simulation over three decades ago, the time scale of reported protein simulations is found to double roughly every 2 years. In terms of length scale, simulations have nearly doubled in size every 6 years. For example, the first molecular dynamics simulation involving bovine pancreatic trypsin inhibitor (BPTI) was carried out for 9.2 picoseconds involving approximately 1082 atoms using a united-atom force field in vacuum [25]. In contrast, it is now fairly routine to simulate models incorporating explicit solvent, periodic boundary conditions, and extended electrostatics with second generation all-atom molecular force fields for 10–100 nanoseconds. As an illustration, lysozyme, a small protein of comparable size to BPTI, has been simulated using a solvated model of explicit waters for 28 nanoseconds involving over 13,000 atoms [26]. Examples including nucleic acid simulation show even greater growth, where a total of 0.6 microseconds of simulation for unique tetranucleotide sequences of DNA containing ~24,000 atoms has been reported [27,28].

It is obvious that the escalation of computing power, resources, and software development has made it easier to create significantly larger and more complex sets of data stemming from molecular dynamics simulations. However, analysis of molecular dynamics trajectories has never been and is currently not trivial. Extracting meaningful information from even the shortest time simulations is an artform requiring solid chemical intuition, physical insight, and technical expertise [29,30]. The increased complexity and size of molecular dynamics trajectories further amplifies an already difficult situation. As such, computational chemists have been searching for new computational tools to mine molecular dynamics data for meaningful information connecting biological function to structure and dynamics. The goal of this review is to demonstrate the need for multivariate

analysis in biophysical studies, present how principal components analysis (PCA) can be implemented in the analysis of molecular dynamics data, and provide insights into the pitfalls and common errors associated with multivariate techniques.

2. MULTIVARIATE METHODS

Systematic variation of a single variable is usually desired in scientific study; however, researchers in the biological, chemical, physical, and social sciences frequently collect measurements on several variables simultaneously. This is especially true for molecular dynamics simulations, where the coordinates and momenta of all atoms are typically sampled every few femtoseconds over millions of time steps. Within the context of molecular dynamics simulations, the challenge is to discover the molecular motion(s) responsible for the phenomena of biochemical interest within the vast range of dynamics ''noise'' [17,18,29,30]. Some progress has been made, where localized molecular motion has been linked to biochemical function as a gateway in acetylcholinesterase [31–33], a hinged-lid in triose phosphate isomerase [34–38], and combined levers and gates in carbonmonoxy myoglobin [39]. A database of more than 120 molecular motions has been reported [40].

A molecular dynamics trajectory is by definition multivariate data, where a large number of variables (atomic positions) are typically found to be interrelated, correlated, or dependent on each other. To decipher these large data sets, multivariate statistical analysis is one approach that is gaining popularity. References that present an organized overview of multivariate methods highlighting their statistical utility and connection between each of the techniques are available [41–43]. There are also excellent sources on individual multivariate methods giving an in-depth mathematical review coupled with illustrative examples and scientific problems suited for such applications [44,45]. For our purposes, multivariate analysis has been applied to molecular dynamics trajectories in two general ways:

(1) *Data reduction or structural simplification*. The goal is to reduce the original large number of dependent variables (atomic coordinates) to a smaller and independent set to explain the phenomena of interest. Data reduction through PCA is unique when applied to molecular dynamics trajectories, since three or less principal components, composed of linear combinations of the original Cartesian coordinates, are typically identified to clarify important biomolecular motions.

(2) *Sorting, classification, and grouping*. The goal is to group or classify objects based upon measured characteristics. In this specific application of multivariate analysis, the dimensionality of the data set remains the same. The data set is simply partitioned into different groups to gain a sense of order or classification. For instance, cluster analysis has been used to identify similar geometric or conformational features from molecular dynamics simulations to further understand complex energy landscapes or design new drugs in pharmaceutical drug design studies.

The method of analysis depends heavily on whether one is interested in interrelationships or in comparisons, and on whether variables are qualitative or quantitative. In many situations, there will not be a single best method of analysis. When applied to molecular dynamics trajectories, the major classifications of multivariate analysis involve PCA [39,46–104], factor analysis [105,106], discriminant function analysis [107], cluster analysis [50,107–122], canonical correlation analysis [123–125], and multidimensional scaling [53,112,113,115,126–130].

A full description for each of these methods is beyond the scope of this review and may be found in other sources [41–45]. There is some overlap between a few of the methods where each technique is generally unique in carrying out either reduction or grouping of multivariate data. However, one of the most commonly applied techniques to molecular dynamics data sets is PCA, which will be the focus of this review.

3. PRINCIPAL COMPONENTS ANALYSIS

3.1. Background

Principal components analysis (PCA) is the simplest of multivariate techniques that is used to reduce or simplify large and complicated sets of data. The PCA procedure was first introduced for only a few variables in 1901 by Karl Pearson [131]. With the advent of computers, PCA was extended as a practical computing method by Hotelling in 1933 for a greater number of variables [132]. Since this time, many variations have been proposed and implemented, such as the essential dynamics method, which has been extensively used and reported in the recent literature. However, the underlying mathematical procedure for essential dynamics remains the same as PCA.

The commonly stated goal of PCA is to reduce the dimensionality of a multivariate data set by taking p interrelated variables, $x_1, x_2, \ldots, x_p$, and finding combinations of these based upon variances to produce a

transformed set of variables, $z_1, z_2, \ldots, z_p$, that are uncorrelated. The indices z_i are called the principal components (PCs). Statistically, the point of PCA is straightforward, but this type of explanation is far from a physical interpretation that would be meaningful to scientists employing such a technique.

It is important first to realize that PCA is predicated on the assumption that the phenomena of interest can be explained by the *variances* and *covariances* between the p variables in the original data set. Unless the number of variables p is small, it is not possible to examine all of the variances or the covariances between the variables manually. PCA overcomes this limitation and transforms the data such that the uncorrelated variables or principal components are ranked by the variance of the data set in a single analysis. In terms of molecular dynamics simulations, PCA ultimately gives a view of the atoms that move anisotropically to maximize the variance.

3.2. Principal components

Before understanding the mathematical process on how PCA is carried out, it is instructive to define the principal components. The first principal component, z_1, is simply a linear combination (dot product) of the original variables $x_1, x_2, \ldots, x_p$, with $\boldsymbol{\alpha}$. Note that the mathematical dot product operator takes two vectors and gives a scalar, or a new variable (principal component) to describe the data.

$$z_1 = \alpha_1^{\mathrm{T}} x = \alpha_{11} x_1 + \alpha_{12} x_2 + \cdots + \alpha_{1p} x_p = \sum_{j=1}^{p} \alpha_{1j} x_j \tag{1}$$

The weights $(\alpha_{11}, \alpha_{12}, \ldots, \alpha_{1p})$ are mathematically determined to maximize the variation of the original data in $\boldsymbol{x}$, subject to the normalization constraint that

$$\alpha_{11}^2 + \alpha_{12}^2 + \cdots + \alpha_{1p}^2 = 1 \tag{2}$$

The constraint is necessary; otherwise, the maximum can simply be increased by increasing any component of $\boldsymbol{\alpha}_j$. To be discussed later, the weights for a particular component are used to interpret and account for the variability in the data. Next, the second principal component, $z_2 = \alpha_2^T \boldsymbol{x}$, is determined having a maximum variance that is uncorrelated with z_1 subject to the same normalization constraint on α_{2p}, and so on, so that the kth principal component, $z_k = \alpha_k{}^T \boldsymbol{x}$, has maximum variance subject to being uncorrelated with $z_1, z_2, \ldots, z_{k\ 1}$. The computed number of principal

components will be same as the number of p original variables. However, in highly correlated data sets, most of the variation from $\boldsymbol{x}$ will be accounted for in a few principal components. In uncorrelated data sets, PCA provides no statistical advantage in treating the data. Obviously, it is desirable to have the value of m much less than the value of p to attain a significant reduction in dimensionality of the data set, where m is the number of principal components necessary to account for the majority of the variation in the data set. The lack of correlation between the principal components is a useful property, since the indices can be interpreted as different ''dimensions'' describing the variation in the data set.

3.3. Covariance matrix

The process of determining the principal components starts with the construction of the $p \times p$ covariance or correlation matrix from a collection of n snapshot structures from molecular dynamics trajectories. The structure matrix $\boldsymbol{x}$ is composed of the Cartesian coordinates for each time stamp that defines each of the rows. The p columns of $\boldsymbol{x}$ are given by the $3N$ Cartesian coordinates for each atom. It is necessary to transform $\boldsymbol{x}$ to remove rotation and translation contamination that does not contribute to the real dynamics of the system. This is accomplished by aligning the structures to a common structure. The reference structure for the alignment process can be an averaged structure, any structure from the trajectory, or an experimental structure. Many techniques have been reported for comparing and overlaying proteins for applications other than for PCA [133–149]. The underlying procedure is essentially the same, where a subset of atoms for the alignment process is selected, and then alignment is carried out using a standard root-mean-square-deviation (*RMSD*) fit on the selected atoms [150]. In studies involving PCA, it is most common to use the alpha carbons or all of the non-hydrogen atoms in the alignment process.

Once $\boldsymbol{x}$ has been aligned, it is possible to compute the covariance matrix elements. The average position $< x_i >$ of the ith atom is computed along the entire trajectory. The convariance between the ith and jth atoms over the collection of n structures can be calculated as shown in equation (3). Each covariance matrix element is determined, as shown in equation (3).

$$c_{ij} = \frac{1}{n}\sum_{k=1}^{n}(x_{ik} - \langle x_i \rangle)(x_{jk} - \langle x_j \rangle) \quad i = 1, 2, \ldots, p \quad j = 1, 2, \ldots, p \qquad (3)$$

The diagonal of the matrix is simply the variance of each coordinate. The covariance is the difference between a variable and its mean multiplied by

the difference of another variable and its mean. Thus, if variable x_i varies largely from its mean, and variable x_j varies largely in the same direction, then the covariance matrix element, c_{ij}, will be large and positive. However, the covariance matrix element for x_i and x_j will be small, if either or both values are close to their corresponding means. With respect to a molecular system, the covariance matrix element between two atoms will be large and positive, if each of those atoms deviate largely from their equilibrium positions and the deviations are in the same direction. Mathematically, the covariance matrix summarizes the covariance between all variable combinations. This matrix is symmetric, so each row and column represents coordinates from the same structures in the same order, i.e. the *k*th row contains the same data points as the *k*th column.

3.4. Index of selectivity

The index of selectivity is simply the set of atoms identified for analysis. The index of selectivity is a modification of the possible values of *i* and *j* in equation (3). It is often assumed in the vast majority of studies utilizing PCA that all atoms should be included in the covariance matrix construction. It is important to realize that selection of all atoms, all non-hydrogen atoms, or all alpha carbons *biases* PCA to extract information involving large-scale global motion. Thus, if localized events are important, and all atoms are selected in the analysis, then the principal components method will likely fail to discover the localized motions, forcing an analysis on motion involving all of the atoms. This problem has been shown for the understanding of the dynamics of carbonmonoxy myoglobin [92]. When all of the non-hydrogen atoms were selected for the PCA, isotropic motion was found, where over 15 dimensions were required to understand the dynamics. However, when smaller and smaller volumes centered about the carbon monoxide ligand were used to select a subset of atoms, the amount of variance was found to be a maximum in two dimensions. Thus, two amino acids, histidine 64 and arginine 45, were found to be responsible for a majority of the anisotropic motion. The dynamics of the two residues were found to explain the spectroscopic A-states of carbonmonoxy myoglobin consistent with available kinetic and mutation data [151–155]. Consequently, the index of selectivity is an important step in the proper use of PCA.

3.5. Eigenanalysis

Analysis using PCA simply involves finding the eigenvectors and eigenvalues of the covariance matrix. The computed eigenvalues from the

covariance matrix are the principal component variances. The eigenvalues are ordered from the largest to the smallest, so that $\lambda_1 \geqslant \lambda_2 \geqslant \ldots \geqslant \lambda_p \geqslant 0$. Specifically, the *k*th eigenvalue, λ_k, indicates the magnitude of the variance of the data in the direction of the corresponding *k*th eigenvector. The resulting eigenvectors provide the coefficients (weights) for the linear combination of observed structures. These eigenvectors are often referred to as the ''loadings'' for the principal components, and referred to as $\boldsymbol{\alpha}_j$ in our previous discussion above. Thus, the linear combination of observed structures, $z_i = \alpha_i{}^T \boldsymbol{x}$, is known as the *i*th principal component.

$$z_i = \alpha_{i1} x_1 + \alpha_{i2} x_2 + \cdots + \alpha_{ip} x_p \tag{4}$$

In protein and nucleic acids, the important data variance can be accounted for by a much smaller number of derived variables (principal components) than the *p* variables from which the analysis begins. For example in the case of nucleic acids, three principal components may account for 85% of the variance in the data [156]. The first two or three principal components often account for enough of the variance that important motions of the protein or nucleic acid can be extracted.

3.6. Scree test dimensionality determination

A key step in PCA is the determination of the number of dimensions to which the data is reduced. This is most easily accomplished by performing the *scree test*, or by creating a *scree plot* [157,158]. This type of plot involves the eigenvalues that are determined in the diagonalization of the covariance matrix. In a scree plot, the *x*-axis is an index of the number of eigenvalues determined. The eigenvalues are ordered from the strongest to weakest. The *y*-axis gives the magnitude of the eigenvalues from the covariance matrix diagonalization. It is customary to scale the eigenvalues such that they sum to unity in order to determine more easily the percent of the variance of the data accounted for the associated eigenvector. To accomplish this, each eigenvalue is divided by the sum of all of the eigenvalues.

To determine the appropriate dimensionality from the resulting analysis, it is necessary to locate the kink in the scree plot, where the variance rapidly falls to a relatively stable value. If the data is highly correlated initially, then the first few dimensions will have large eigenvalues, which indicates that a great amount of variance is described in those dimensions. The variance should drop rapidly and form a relatively flat plateau. The correct dimensionality is typically the dimension prior to the eigenvalue reaching the plateau. The interpretation is such that adding the extra dimension does not

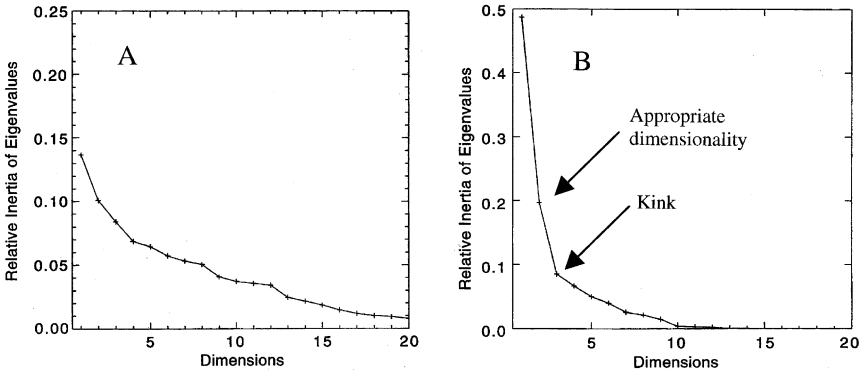

Fig. 1. Scree plots using indices of selectivity that include (a) all heavy atoms and (b) His64, the CO ligand and the heme.

result in any appreciable gain in information (variance) on the system, as compared to the complexity of adding an additional dimension.

Two example scree plots from our earlier work on carbonmonoxy myoglobin are given below [92]. From both plots, it is possible to identify the associated problems with the improper use of the index of selectivity, as commonly assumed (Fig. 1).

When all atoms are used in the PCA, a well-defined kink is never realized. The relative inertia monotonically decreases as the dimension increases. A scree plot with this type of signature indicates that the variance of the molecular system cannot be consolidated under a few dimensions. In using such an index of selectivity, it would be impossible to determine the correct dimensionality for further analysis. However, in Fig. 1b, where a subset of atoms is used to define the index of selectivity, it is clear that a significant portion of the system's variance is captured in the first two dimensions. In fact, approximately 70% of the information is found in the first two principal components. In this specific case, the third dimension delivers additional 10% of information; however, no useful data was found upon examination. The two plots illustrate the problem associated with assuming that all of the atoms should be used in the index of selectivity.

3.7. Visualization

The next step is visualization of the data using the dimensionality determined from the previous step. Projections of the original structures onto the weights (eigenvectors) of the associated principal components are plotted against each other as the scree plot dictates. Consequently, if *M* structures are collected and used in the covariance matrix construction,

then *M* data points will be realized on the plots. The principal components plots give information on the similarity of the *M* structures used to form the covariance matrix. As an example, the two-dimensional plot corresponding to 1045 structures and scree plot in Fig. 1b is given below.

Each point on the plot corresponds to a structure and its relation to all other structures in the dimension(s) plotted. If two points are close to each other on the plot, then those structures are similar. If two points on such a plot are far from each other, then they are dissimilar in some fashion. It is clear that this specific plot yields three general basins of structural similarity. This behavior is consistent with the current ideas of energy landscapes, where multiple minima are clustered into regions that are separated by higher energy barriers [159]. It is at this juncture that the origin of dissimilarity between the three energy basins cannot be derived from the principal components plots alone. All that is known is that the structures are different given the index of selectivity utilized. In this specific instance, the difference was determined to be a result of the structural change of histidine 64, since that was the primary constituent of the index of selectivity. More traditional methods such as constructing the *RMSD* between the two structures from the different energy basins as a function of its sequence can usually pinpoint the molecular reasons of conformational differences. It is through the pair wise comparison of structures from the different basins that a molecular interpretation may be formulated in describing the different conformations sampled by the molecular system. A major result of the Schulze and Evanseck study was that histidine 64 moved in from the solvent to the ligand by $\sim$10 Å on a timescale consistent with experiment [151], as shown in Fig. 2b.

When more than three dimensions are indicated by the scree plot, it is possible to examine the variance at the higher dimensions. As an example, when the first and second principal components are determined (using the scree test) to describe large amounts of variance, and when the third and fourth are very close together in magnitude, two separate plots may be created to help characterize molecular motion. The first plot may have the first, second, and third principal components as the axes, and the second plot would use first, second, and fourth as the axes.

4. ESSENTIAL DYNAMICS

4.1. Background

Certain types of internal motions allow proteins to perform their biological functions. These motions may enable the binding of substrates, adaptation

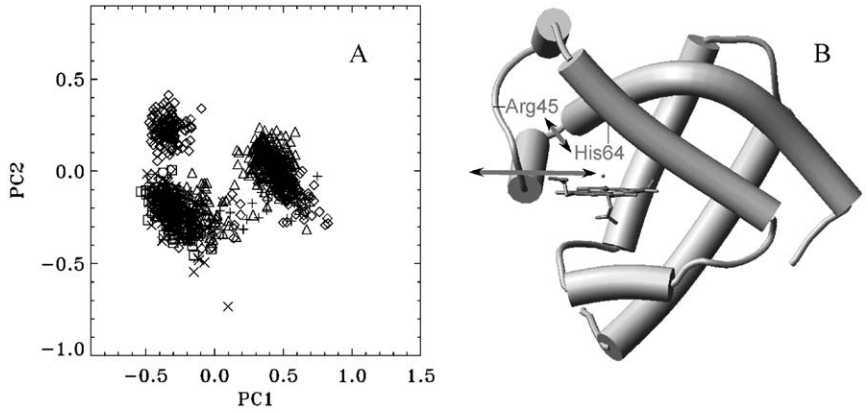

Fig. 2. (a) The two dimensional principal component plot of carbonmonoxy myoglobin using the coordinates of His64, the ligand and heme. The different symbols indicate different starting conditions of the multiple short time trajectories used to form the ensemble of structures. (b) Vectors showing the conformational change extracted from Fig. 2a.

to various environments, or conformational changes that allow binding of a substrate at another site on the protein (allosteric effects). The internal motions may be subtle and can involve complicated correlations between atomic motions, and thus the challenge presented is to identify these motions, determine how they relate to protein function, and to separate the complicated dynamics from the essential degrees of freedom [102,160,161]. Amadei and coworkers first developed the method known as essential dynamics in order to separate the concerted structural rearrangement from irrelevant motions [161]. Their method is based on the hypothesis that by using PCA, atomic positional fluctuations can be used to separate a protein's conformational space into two subspaces: an "essential" subspace which contains only a few degrees of freedom that describe the motions relevant for protein function (e.g., opening and closing and hinge bending motions) and the remaining subspace ("constrained subspace") that describes the irrelevant local fluctuations of the protein. This group used lysozyme as their test case of the essential dynamics method, and they concluded that the essential dynamics of most proteins can be described in a subspace of only a few degrees of freedom, while all other degrees of freedom represent much less important and mostly independent fluctuations of the molecule.

The essential dynamics method involves the use of a covariance matrix constructed from structures sampled throughout a molecular dynamics simulation. By diagonalization of a covariance matrix of the atomic

coordinates of the system, the motions of a structure that are responsible for the most variance in atomic position are targeted. The essential subspace is determined by ranking the eigenvalues elucidated by PCA, of the covariance matrix from the molecular dynamics trajectory. The mathematical equivalence between PCA and essential dynamics has been noted before [88,162–168] and has described within this document.

4.2. Applications to protein systems

Many authors of protein simulation studies have used the essential dynamics method in order to identify important molecular vibrations to understand more about large correlated protein motions and how they are critical to biological function. PCA has been used in a wide array of applications ranging from crystallographic and NMR structure ensembles [63,64,73,80,169–177], protein and peptide folding/unfolding [47,54,66,67,99,167,178–184], structural determinants of transmembrane proteins and channels [49,51,55,69,90,101,185–191], large-scale domain motion [58,77,98,104,192–199], locally accessible conformational substates [52,60,92,96,97,103,200], correlated and functional motion [39,56,57,61,71,84,163,166,168,201–215], dynamic effects from mutations and domain swapping [216–219], mutation impact upon binding [220–223], connection between structural similarity and dynamic behavior [87,89,164,224], ligand binding and migration [53,74,82,225–233], conformations of small molecules [72,79,91], protonation effects on dynamics [234], liquid behavior and spectroscopy [48,75,76], testing and development of methodology [46,59,62,65,78,83,85,86,235], protein docking algorithms [68,70,236–238], homology modeling [100], and atomic and molecular properties [50,95].

4.3. Applications to nucleic acids

PCA has been shown to be a powerful tool in evaluation of DNA flexibility in molecular dynamics simulations [162,239–241]. This technique has been employed to examine nucleic acid flexibility [239,240,242], flexibility of hybrid nucleic acids [243], flexibility of DNA in the crystal environment [240], behavior of A-tract DNA [93], electrostatic interactions of nucleic acids [83,156], sequence effects [27,28], DNA containing chemical modifications [81,244–246], broken strand DNA [247], base flipping [248], and nucleic acid mispairs [242]. The potential energy surface of nucleic acid conformational changes have also been investigated using PCA [156]. As new techniques in molecular dynamics simulations emerge, PCA has been used to evaluate the quality of simulations [249–251].

5. RELATED METHODS

5.1. Independent component analysis

Independent component analysis (ICA) is a multivariate technique that is used to separate independent variables in a data set [252,253]. Unlike PCA, ICA is not typically used as a dimensionality reduction technique. In ICA, the data must be fit to a model (not necessarily a linear model) [254] in which the derived variables are as statistically independent as possible. In chemical applications, it is generally favorable if the number of derived variables (principal components, latent variables) is much smaller than the number of original observations collected. Therefore, a data reduction technique such as PCA may be performed before standard ICA is carried out [252,255]. Thus, the focus of ICA is on the subspace accounting for the most variance in the data set when the analysis begins [252]. Westad and Kermit investigated validation methods for ICA and found that cross-validation was a valuable tool for determining the number of principal components to use from the preliminary reduction step and the number of ICs to extract in the actual ICA [256]. Yadava and Chaudhary applied ICA to determine analyte solvation parameters on polymer-coated surface acoustic wave vapor sensors [255]. ICA was employed because PCA did not yield derived variables that were interpretable for this particular type of experiment. ICA has been used in analysis of spectroscopic data [257,258]. Medical image processing is also an area in which ICA has been used [259]. Other uses of ICA include analysis of natural systems such as seismological and atmospheric data [260] and atmospheric aerosol content [261].

5.2. Singular value decomposition

The singular value decomposition (SVD) technique was established by several mathematicians who worked independently to develop the theory leading to the efficient diagonalization of a matrix [262]. Although SVD has many uses, it is commonly used to extract eigenvalues from a symmetric matrix [263]. As such, the technique has been used in PCA [94,264,265]. SVD may be used as a tool to execute PCA on a variety of systems including NMR spectroscopy [194,266] and X-ray photoelectron spectromicroscopy [265]. Andrews and coworkers used the SVD algorithm to perform PCA on myoglobin. In their work, SVD was chosen because it is computationally efficient. SVD was carried out on the internal coordinates of the myoglobin, which gave similar results as the SVD of the Cartesian coordinates [94]. Tomfohr and coworkers used SVD to diagonalize a ma-

trix for dimensionality reduction of gene expression data [264]. SVD may also be used to carry out Gaussian network model analysis [267].

6. LIMITATIONS AND COMMON ERRORS

Multivariate techniques can be very powerful in data analysis. However, there are only a few papers that critically examine the possible weaknesses of multivariate analyses [102,268]. When using these statistical tools on molecular dynamics simulations, one should realize that there exist potential sources of error that could bias the analysis and provide misleading or wrong interpretations of the data.

The first and most important source of error deals with the well-documented sampling issues with molecular dynamics simulations [8,9,11,13,39]. The goal of applying PCA to molecular dynamics trajectories is to extract and understand the dynamics of the system. Consequently, if the trajectory samples only a portion of available structures from the true ensemble, then PCA will extract and provide information on the incomplete representation of phase space. Multivariate analysis will not create data to correct problems with the generation of the original data set.

Secondly, the index of selectivity is crucial to a successful PCA, which is often overlooked in a majority of studies utilizing dimensionality reduction and molecular dynamics. Care needs to be exercised in atom selection, where all atoms, all nonhydrogen atoms, or all alpha carbons of proteins are typically used for analysis. Selection based upon all atoms is correct, as long as low-frequency, large-scale motions are desired. However, it should be clear that molecular motion need not be large scale. As mentioned before, many well-understood examples show that local-motion is connected with function, as gateways [31–33], hinged-lids [34–38], and combined levers and gates [39]. Therefore, important motions could be localized and only a subset of the atoms is needed within the range of molecular motions. In carbonmonoxy myoglobin, it was necessary to modify the index of selectivity, based upon previous knowledge of the binding site, in order to discover the local motion responsible for the spectroscopic A-states [39]. Indices of selectivity can bias multivariate analysis, where it is necessary to have a course idea of the type of dynamics of interest, i.e., local or global motion, in molecular dynamics simulations.

Lastly, when working with PCA, it is essential to bear in mind that the major assumption is that the sources of largest variance are of importance to the problem being addressed. However, caution needs to be exercised

in mixed data sets that involve more than the atomic coordinates from molecular dynamics trajectories. For example, differences in the units could be involved, where the original data may be composed of differently measured characteristics. For example, the variation in angstroms in atomic position is obviously different than the variation of pH or temperature. Even when the same units are used, it is plausible that one measured quantity may have a completely different range of behavior compared to another. Consider the variation in covalent bond length versus the variation in intermolecular hydrogen bonding. When variables with large variance are compared with variables of small variance, those with larger associated variance will be weighted more heavily in construction of the principal components. This weighting is simply due to the fact that the goal in constructing the principal components is to maximize variance. In cases with variables with widely ranging variance, using a covariance matrix of standardized variables, or correlation matrix to determine the principal components may help to alleviate this issue [45].

7. CONCLUSION

The continued advances in readily available computer power coupled with the desire to explore dynamics at longer time scales means that the magnitude and complexity of accessible dynamics data will keep growing. By necessity, methods to reduce the size of this data will continue to be valued by computational chemists. In this review, we have sought to highlight the utility of PCA to reduce the complexity of variables describing the dynamics data. PCA and the mathematically identical essential dynamics, have proved useful in the detection of important motions in biomolecules ranging from proteins to nucleic acids. Provided that appropriate care is taken with the use of these methods, computational chemists should find PCA useful in managing large, complex data sets and discovering molecular motions that are biochemically relevant.

ACKNOWLEDGMENTS

This work was funded in part by the National Science Foundation (CHE-0321147, CHE-0354052, AAB/PSC CHE-030008P), Department of Education (P116Z040100 and P116Z050331), SGI and Clarix Corporations, the National Institutes of Health (GM069549-01), and the Center for Computational Sciences at Duquesne University.

REFERENCES

[1] M. Karplus and J. Kuriyan, Molecular dynamics and protein function, *Proc. Natl. Acad. Sci.*, 2005, **102**, 6679–6685.
[2] T. Hansson, C. Oostenbrink and W. F. van Gunsteren, Molecular dynamics simulations, *Curr. Opin. Struc. Biol.*, 2002, **12**, 190–196.
[3] M. Karplus, Molecular dynamics simulations of biomolecules, *Acc. Chem. Res.*, 2002, **35**, 321–323.
[4] M. Karplus and J. A. McCammon, Molecular dynamics simulations of biomolecules, *Nat. Struct. Biol.*, 2002, **9**, 646–652.
[5] W. Wang, O. Donini, C. M. Reyes and P. A. Kollman, Biomolecular simulations: Recent developments in force fields, *simulations of enzyme catalysis, protein-ligand, protein-protein, and protein-nucleic acid noncovalent interactions. Annu. Rev. Biophys. Biomol. Struct.*, 2001, **30**, 211–243.
[6] M. Karplus and G. A. Petsko, Molecular dynamics simulations in biology, *Nature*, 1990, **347**, 631–639.
[7] M. Karplus, Molecular dynamics simulations of proteins, *Phys. Today*, 1987, **40**, 68–70.
[8] X. Cheng, G. Cui, V. Hornak and C. Simmerling, Modified replica exchange simulation methods for local structure refinement, *J. Phys. Chem. B*, 2005, **109**, 8220–8230.
[9] A. E. Loccisano, O. Acevedo, J. DeChancie, B. G. Schulze and J. D. Evanseck, Enhanced sampling by multiple molecular dynamics trajectories: carbonmonoxy myoglobin 10 microsecond $A_0 > A_{1\text{-}3}$ transition from ten 400 picosecond simulations, *J. Mol. Graph. Model*, 2004, **22**, 369–376.
[10] P. Minary, M. E. Tuckerman and G. T. Martyna, Long time molecular dynamics for enhanced conformational sampling in biomolecular systems, *Phys. Rev. Lett.*, 2004, **93**, 1520201/1–1520201/4.
[11] I. Andricioaei, A. R. Dinner and M. Karplus, Self-guided enhanced sampling methods for thermodynamic averages, *J. Chem. Phys.*, 2003, **118**, 1074–1084.
[12] T. Schlick, *Molecular Modeling and Simulation*, Springer, New York, 2002.
[13] Z. Zhu, M. E. Tuckerman, S. O. Samuelson and G. T. Martyna, Using novel variable transformations to enhance conformational sampling in molecular dynamics, *Phys. Rev. Lett.*, 2002, **88**, 100201/1–100201/4.
[14] H. Grubmuller, Predicting slow structural transitions in macromolecular systems: Conformational Flooding, *Phys. Rev. E*, 1995, **52**, 2893–2906.
[15] G. Bhanota, D. Chen, A. Gara and P. Vranas, The BlueGene/L supercomputer, *Nucl. Phys. B (Proc. Suppl.)*, 2003, **119**, 114–121.
[16] F. Bodin, P. Boucaud, N. Cabibbo, G. Cascino, F. Calvayrac, M. Della Morte, A. Del Re, R. De Pietri, P. Deriso and F. Di Carlo, APE computers – past, present and future, *Comput. Phys. Commun.*, 2002, **147**, 402–409.
[17] A. H. Zewail, Femtochemistry, Atomic-scale dynamics of the chemical bond using ultrafast lasers Nobel lecture. In *Les Prix Nobel* (ed. T. Frangsmyr), Almqvist and Wiksell International, Stockholm, 2000, pp. 110–203.
[18] R. M. Hochstrasser, Ultrafast spectroscopy of protein dynamics, *J. Chem. Educ.*, 1998, **75**, 559–564.
[19] V. Reat, H. Patzelt, M. Ferrand, C. Pfister, D. Oesterhelt and G. Zaccai, Dynamics of different functional parts of bacteriorhodopsin: H-2 H labeling and neutron scattering, *Proc. Natl. Acad. Sci.*, 1998, **95**, 4970–4975.
[20] M. Ben-Nun, J. Cao and K. R. Wilson, Ultrafast X-ray and electron diffraction: Theoretical considerations, *J. Phys. Chem. A*, 1997, **101**, 8743–8761.
[21] E. Chen, R. A. Goldbeck and D. S. Kliger, Nanosecond time-resolved spectroscopy of biomolecular processes, *Annu. Rev. Biophys. Biomol. Struct.*, 1997, **26**, 327–355.

[22] T. Schlick, E. Barth and M. Mandziuk, Biomolecular dynamics at long timesteps: Bridging the timescale gap between simulation and experimentation, *Annu. Rev. Biophys. Biomol. Struct.*, 1997, **26**, 181–222.
[23] Y. Duan and P. A. Kollman, Pathways to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution, *Science*, 1998, **282**, 740–744.
[24] G. E. Moore, Cramming more components onto integrated circuits, *Electronics*, 1965, **38**, 114–117.
[25] J. A. McCammon, B. R. Gelin and M. Karplus, Dynamics of folded proteins, *Nature*, 1977, **267**, 585–590.
[26] M. Marchi, F. Sterpone and M. Ceccarelli, Water rotational relaxation and diffusion in hydrated lysozyme, *J. Am. Chem. Soc.*, 2002, **124**, 6787–6791.
[27] S. B. Dixit, D. L. Beveridge, D. A. Case, T. E. Cheatham III, E. Giudice, F. Lankas, R. Lavery, J. H. Maddocks, R. Osman, H. Sklenar, K. M. Thayer and P. Varnai, Molecular dynamics simulations of the 136 unique tetranucleotide sequences of DNA oligonucleotides. II: Sequence context effects on the dynamical structures of the 10 unique dinucleotide steps, *Biophys. J.*, 2005, **89**, 3721–3740.
[28] D. L. Beveridge, G. Barreiro, K. S. Byun, D. A. Case, T. E. Cheatham III, S. B. Dixit, E. Giudice, F. Lankas, R. Lavery, J. H. Maddocks, R. Osman, E. Seibert, H. Sklenar, G. Stoll, K. M. Thayer, P. Varnai and M. A. Young, Molecular dynamics simulations of the 136 unique tetranucleotide sequences of DNA oligonucleotides, I. Research design and results on d(CpG) steps, *Biophys. J.*, 2004, **87**, 3799–3813.
[29] C. L. Brooks, III, M. Karplus and B. M. Pettitt, *Proteins: A Theoretical Perspective of Dynamics, Structure, and Thermodynamics*, Wiley, New York, 1988.
[30] J. A. McCammon and S. C. Harvey, *Dynamics of Proteins and Nucleic Acids*, Cambridge University Press, Cambridge, 1988.
[31] T. Y. Shen, T. Kaihsu and J. A. McCammon, Statistical analysis of the fractal gating motions of the enzyme acetylcholinesterase, *Phys. Rev. E*, 2001, **63**, 041902/1–041902/6.
[32] N. A. Baker and J. A. McCammon, Non-Boltzmann rate distributions in stochastically gated reactions, *J. Phys. Chem. B*, 1999, **103**, 615–617.
[33] H.-X. Zhou, S. T. Wlodek and J. A. McCammon, Conformation gating as a mechanism for enzyme specificity, *Proc. Natl. Acad. Sci.*, 1998, **95**, 9280–9283.
[34] J. Sun and N. S. Sampson, Understanding protein lids: Kinetic analysis of active hinge mutants in triosephosphate isomerase, *Biochemistry*, 1999, **38**, 11474–11481.
[35] P. Derreumaux and T. Schlick, The loop opening/closing motion of the enzyme triosephosphate isomerase, *Biophys. J.*, 1998, **74**, 72–81.
[36] K. Yuksel, A. Sun, R. Gracy and K. Schnackerz, The hinged lid of yeast triosephosphate isomerase. Determination of the energy barrier between the two conformations, *J. Biol. Chem.*, 1994, **269**, 5005–5008.
[37] N. S. Sampson and J. R. Knowles, Segmental motion in catalysis: Investigation of a hydrogen bond critical for loop closure in the reaction of triosephosphate isomerase, *Biochemistry*, 1992, **31**, 8488–8494.
[38] D. Joseph, G. A. Petsko and M. Karplus, Anatomy of a conformational change: Hinged ''lid'' motion of the triosephosphate isomerase loop, *Science*, 1990, **249**, 1425–1428.
[39] B. G. Schulze, H. Grubmuller and J. D. Evanseck, Functional significance of hierarchical tiers in carbonmonoxy myoglobin: Conformational substates and transitions studied by conformational flooding simulations, *J. Am. Chem. Soc.*, 2000, **122**, 8700–8711.
[40] M. Gerstein and W. Krebs, A database of macromolecular motions, *Nucleic Acids Res*, 1998, **26**, 4280–4290.
[41] L. G. Grimm and P. R. Yarnold, *Reading and Understanding Multivariate Statistics*, American Psychological Association, Washington, DC, 1998.

[42] B. F. J. Manly, *Multivariate Statistical Methods: A Primer*, Chapman and Hall, London, 1994.
[43] R. A. Johnson and D. W. Wichern, *Applied Multivariate Statistical Analysis*, Prentice Hall, Upper Saddle River, 1992.
[44] I. T. Jolliffe, *In Principal Component Analysis*, 2nd ed, Springer, New York, 2002.
[45] G. H. Dunteman, In: M.S. Lewis-Beck (Ed.), Principal Components Analysis, 1st Ed., vol. 69, Sage, Newbury Park, 1989, p. 96–97.
[46] C. P. Barrett and M. E. M. Noble, Dynamite extended: Two new services to simplify protein dynamic analysis, *Bioinformatics*, 2005, **21**, 3174–3175.
[47] C. Chen, Y. Xiao and L. Zhang, A directed essential dynamics simulation of peptide folding, *Biophys. J.*, 2005, **88**, 3276–3285.
[48] M. D'Abramo, M. D'Alessandro, A. Di Nola, D. Roccatano and A. Amadei, Characterization of liquid behavior by means of local density fluctuations, *J. Mol. Liq.*, 2005, **117**, 17–21.
[49] S. Haider, A. Grottesi, B. A. Hall, F. M. Ashcroft and M. S. P. Sansom, Conformational dynamics of the ligand-binding domain of inward rectifier K channels as revealed by molecular dynamics simulations: Toward an understanding of Kir channel gating, *Biophys. J.*, 2005, **88**, 3310–3320.
[50] O. Horovitz and C. Sarbu, Characterization and classification of lanthanides by multivariate analysis methods, *J. Chem. Ed.*, 2005, **82**, 473–483.
[51] A. Hung, K. Tai and M. S. P. Sansom, Molecular dynamics simulation of the M2 helices within the nicotinic acetylcholine receptor transmembrane domain: Structure and collective motions, *Biophys. J.*, 2005, **88**, 3321–3333.
[52] A. Leo-Macias, P. Lopez-Romero, D. Lupyan, D. Zerbino and A. R. Ortiz, An analysis of core deformations in protein superfamilies, *Biophys. J.*, 2005, **88**, 1291–1299.
[53] Y. Li, Z. Zhou and C. B. Post, Dissociation of an antiviral compound from the internal pocket of human rhinovirus 14 capsid, *Proc. Natl. Acad. Sci.*, 2005, **102**, 7529–7534.
[54] J. T. MacDonald, A. G. Purkiss, M. A. Smith, P. Evans, J. M. Goodfellow and C. Slingsby, Unfolding crystallins: The destabilizing role of a β-hairpin cysteine in βB2-crystallin by simulation and experiment, *Protein Sci*, 2005, **14**, 1282–1292.
[55] S. Oyama, Jr., P. Pristovsek, L. Franzoni, A. Pertinhez Thelma, E. Schinina, C. Lucke, H. Ruterjans, C. Arantes Eliane and A. Spisni, Probing the pH-dependent structural features of α-KTx12.1, a potassium channel blocker from the scorpion Tityus serrulatus, *Protein Sci*, 2005, **14**, 1025–1038.
[56] P. W. Pan, R. J. Dickson, H. L. Gordon, S. M. Rothstein and S. Tanaka, Functionally relevant protein motions: Extracting basin-specific collective coordinates from molecular dynamics trajectories, *J. Chem. Phys.*, 2005, **122**, 034904.
[57] G. R. Smith, M. J. Sternberg and P. A. Bates, The relationship between the flexibility of proteins and their conformational states on forming protein-protein complexes with an application to protein-protein docking, *J. Mol. Biol.*, 2005, **347**, 1077–1101.
[58] Z. Zhou, M. Madrid, J. D. Evanseck and J. D. Madura, Effect of a bound nonnucleoside RT inhibitor on the dynamics of wild-type and mutant HIV-1 reverse transcriptase, *J. Am. Chem. Soc.*, 2005, **127**, 17253–17260.
[59] L. Afzelius, F. Raubacher, A. Karlen, F. S. Jorgensen, T. B. Andersson, C. M. Masimirembwa and I. Zamora, Structural analysis of CYP2C9 and CYP2C5 and an evaluation of commonly used molecular modeling techniques, *Drug Metab. Dispos.*, 2004, **32**, 1218–1229.
[60] B. Alakent, P. Doruker and M. C. Camurdan, Application of time series analysis on molecular dynamics simulations of proteins: A study of different conformational spaces by principal component analysis, *J. Chem. Phys.*, 2004, **121**, 4759–4769.
[61] B. Alakent, P. Doruker and M. C. Camurdan, Time series analysis of collective motions in proteins, *J. Chem. Phys.*, 2004, **120**, 1072–1088.
[62] C. P. Barrett, B. A. Hall and M. E. M. Noble, Dynamite: A simple way to gain insight into protein motions, *Acta Cryst. D.*, 2004, **60**, 2280–2287.

[63] F. Corzana, S. Motawia Mohammed, C. Herve du Penhoat, F. van den Berg, A. Blennow, S. Perez and B. Engelsen Soren, Hydration of the amylopectin branch point. Evidence of restricted conformational diversity of the α-(1 $>$6) linkage, *J. Am. Chem. Soc.*, 2004, **126**, 13144–13155.
[64] E. G. Emberly, R. Mukhopadhyay, C. Tang and N. S. Wingreen, Flexibility of β-sheets: Principal component analysis of database protein structures, *Proteins*, 2004, **55**, 91–98.
[65] D. Flock, I. Daidone and A. Di Nola, A molecular dynamics study of acylphosphatase in aggregation-promoting conditions: The influence of trifluoroethanol/water solvent, *Biopolymers*, 2004, **75**, 491–496.
[66] N. J. Marianayagam and S. E. Jackson, The folding pathway of ubiquitin from all-atom molecular dynamics simulations, *Biophys. Chem.*, 2004, **111**, 159–171.
[67] A. Palazoglu, A. Gursoy, Y. Arkun and B. Erman, Folding dynamics of proteins from denatured to native state: Principal component analysis, *J. Comp. Biol.*, 2004, **11**, 1149–1168.
[68] R. Tatsumi, Y. Fukunishi and H. K. Nakamura, A hybrid method of molecular dynamics and harmonic dynamics for docking of flexible ligand to flexible receptor, *J. Comp. Chem.*, 2004, **25**, 1995–2005.
[69] Y. S. Watanabe, Y. Fukunishi and H. K. Nakamura, Modelling of third cytoplasmic loop of bovine rhodopsin by multicanonical molecular dynamics, *J. Mol. Graph. Model*, 2004, **23**, 59–68.
[70] M. Zacharias, Rapid protein–ligand docking using soft modes from molecular dynamics simulations to account for protein deformability: Binding of FK506 to FKBP, *Proteins*, 2004, **54**, 759–767.
[71] G. Chillemi, P. Fiorani, P. Benedetti and A. Desideri, Protein concerted motions in the DNA-human topoisomerase I complex, *Nucleic Acids Res.*, 2003, **31**, 1525–1535.
[72] X. Fradera, M. Marquez, B. D. Smith, M. Orozco and F. J. Luque, Molecular dynamics study of [2]rotaxanes: Influence of solvation and cation on co-conformation, *J. Org. Chem.*, 2003, **68**, 4663–4673.
[73] J.-C. Hus, W. Peti, C. Griesinger and R. Bruschweiler, Self-consistency analysis of dipolar couplings in multiple alignments of ubiquitin, *J. Am. Chem. Soc.*, 2003, **125**, 5596–5597.
[74] A. Nijnik, R. Mott, D. P. Kwiatkowski and I. R. Udalova, Comparing the fine specificity of DNA binding by NF-κB p50 and p52 using principal coordinates analysis, *Nucleic Acids Res.*, 2003, **31**, 1497–1501.
[75] R. A. Wheeler and H. Dong, Optimal spectrum estimation in statistical mechanics, *ChemPhysChem*, 2003, **4**, 1227–1230.
[76] R. A. Wheeler, H. Dong and S. E. Boesch, Quasiharmonic vibrations of water, water dimer, and liquid water from principal component analysis of quantum and QM/MM trajectories, *ChemPhysChem*, 2003, **4**, 382–384.
[77] N. P. Barton, C. S. Verma and L. S. D. Caves, Inherent flexibility of calmodulin domains: A normal-mode analysis study, *J. Phys. Chem. B.*, 2002, **106**, 11036–11040.
[78] L. S. D. Caves and C. S. Verma, Congruent qualitative behavior of complete and reconstructed phase space trajectories from biomolecular dynamics simulation, *Proteins Struct. Funct. Genet.*, 2002, **47**, 25–30.
[79] M. D'Alessandro, A. Tenenbaum and A. Amadei, Coherent dynamics in a butane molecule, *Phys. Rev. E*, 2002, **66**, 020901/1–020901/4.
[80] R. Dvorsky, V. Hornak, J. Sevcik, G. P. Tyrrell, L. S. D. Caves and C. S. Verma, Dynamics of RNase Sa: A simulation perspective complementary to NMR/X-ray, *J. Phys. Chem. B*, 2002, **106**, 6038–6048.
[81] H. Ishida, Molecular dynamics simulation of 7,8-dihydro-8-oxoguanine DNA, *J. Biomol. Struct. Dyn.*, 2002, **19**, 839–851.

[82] M. J. Millan, L. Maiofiss, D. Cussac, V. Audinot, J. A. Boutin and A. Newman-Tancredi, Differential actions of anti-Parkinson agents at multiple classes of monoaminergic receptor. 1. A multivariate analysis of the binding profiles of 14 drugs at 21 native and cloned human receptor subtypes, *J. Pharm. Exp. Ther.*, 2002, **303**, 791–804.
[83] M. Nina and T. Simonson, Molecular dynamics of the tRNA Ala acceptor stem: Comparison between continuum reaction field and Particle-Mesh Ewald electrostatic treatments, *J. Phys. Chem. B.*, 2002, **106**, 3696–3705.
[84] J. T. A. Saarala, K. Tuppurainen, M. Perakyla, H. Santa and Laatikainen, Correlative motions and memory effects in molecular dynamics simulations of molecules: Principal components and rescaled range analysis suggest that the motions of native BPTI are more correlated than those of its mutants, *Biophys. Chem.*, 2002, **95**, 49–57.
[85] N. Ota and D. A. Agard, Enzyme specificity under dynamic control II. Principal components analysis of a-lytic protease using global and local solvent boundary conditions,, *Protein Sci.*, 2001, **10**, 1403–1414.
[86] R. Dvorsky, J. Sevcik, L. S. D. Caves, R. E. Hubbard and C. S. Verma, Temperature effects on protein motions: A molecular dynamics study of RNase-Sa, *J. Phys. Chem. B*, 2000, **104**, 10387–10397.
[87] A. Giuliani, R. Benigni, P. Sirabella, J. P. Zbilut and A. Colosimo, Nonlinear methods in the analysis of protein sequences: A case study in rubredoxins, *Biophys. J.*, 2000, **78**, 136–148.
[88] B. Hess, Similarities between principal components of protein dynamics and random diffusion, *Phys. Rev. E*, 2000, **62**, 8438–8448.
[89] M. A. Ceruso, A. Amadei and A. Di Nola, Mechanics and dynamics of B1 domain of Protein G: Role of packing and surface hydrophobic residues, *Protein Sci.*, 1999, **8**, 147–160.
[90] J. M. Koshi and W. J. Bruno, Major structural determinants of transmembrane proteins identified by principal components analysis, *Proteins*, 1999, **34**, 333–340.
[91] H. Lanig, M. Gottschalk, S. Schneider and T. W. Clark, Conformational analysis of tetracycline using molecular mechanical and semiempirical MO-calculations, *J. Mol. Mod.*, 1999, **5**, 46–62.
[92] B. G. Schulze and J. D. Evanseck, Cooperative role of Arg45 and His64 in the spectroscopic A3 state of carbonmonoxy myoglobin: Molecular dynamics simulations, multivariate anlaysis and quantum mechanical computations, *J. Am. Chem. Soc.*, 1999, **121**, 6444–6454.
[93] E. C. Sherer, S. A. Harris, R. Soliva, M. Orozco and C. A. Laughton, Molecular dynamics studies of DNA A-tract structure and flexibility, *J. Am. Chem. Soc.*, 1999, **121**, 5981–5991.
[94] B. K. Andrews, T. Romo, J. B. Clarage, B. M. Pettitt and G. N. Phillips, Jr., Characterizing global substates of myoglobin, *Structure*, 1998, **6**, 587–594.
[95] E. Bolzacchini, V. Consonni, R. Lucini, M. Orlandi and B. Rindone, High-performance size-exclusion chromatographic behavior of substituted benzoylpoly L-lysines by principal component analysis and molecular dynamics simulations, *J. Chromatogr. A*, 1998, **813**, 255–265.
[96] L. S. D. Caves, J. D. Evanseck and M. Karplus, Locally accessible conformations of proteins: Multiple molecular dynamics simulations of crambin, *Protein Sci.*, 1998, **7**, 649–666.
[97] R. Laatikainen, J. T. A. Saarala, K. Tuppurainen and T. Hassinen, Internal motions of native lysozyme are more organized than those of mutants: A principal component analysis of molecular dynamics data, *Biophys. Chem.*, 1998, **73**, 1–5.
[98] S. Hayward, A. Kitao and H. J. C. Berendsen, Model-free methods of analyzing domain motions in proteins from simulation: A comparison of normal mode analysis and molecular dynamics simulation of lysozyme, *Proteins*, 1997, **27**, 425–437.

[99] T. Lazaridis, I. Lee and M. Karplus, Dynamics and unfolding pathways of a hyperthermophilic and a mesophilic rubredoxin, *Protein Sci*, 1997, **6**, 2589–2605.
[100] K. Ogata and H. Umeyama, Prediction of protein side-chain conformations by principal component analysis for fixed main-chain atoms, *Protein Eng.*, 1997, **10**, 353–359.
[101] S. T. Wlodek, T. W. Clark, L. R. Scott and J. A. McCammon, Molecular dynamics of acetylcholinase dimer complexed wtih tacrine, *J. Am. Chem. Soc.*, 1997, **119**, 9513–9522.
[102] M. A. Balsera, W. Wriggers, Y. Oono and K. Schulten, Principal component analysis and long time protein dynamics, *J. Phys. Chem.*, 1996, **100**, 2567–2572.
[103] S. Hayward, A. Kitao and N. Go, Harmonic and anharmonic aspects in the dynamics of BPTI: A normal mode analysis and principal component analysis, *Protein Sci.*, 1994, **3**, 936–943.
[104] A. E. Garcia, Large-amplitude nonlinear motions in proteins, *Phys. Rev. Lett.*, 1992, **68**, 2696–2699.
[105] M. Kronen, H. Gorls, H.-H. Nguyen, S. Reissman, M. Bohl, J. Suhnel and U. Grafe, Crystal structure and conformational analysis of ampullosporin A, *J. Pept. Sci.*, 2003, **9**, 729–744.
[106] J. Hanus, I. Barvik, K. Ruszovz-Chmelova, J. Stepanek, P.-Y. Turpin and J. Bok, I. Rosenberg and M. Petrova-Endova, -CH2-lengthening of the internucleotide linkage in the ApA dimer can improve its conformational compatibility with its natural polynucleotide counterpart, *Nucleic Acids Res.*, 2001, **29**, 5182–5194.
[107] Y. K. Reshetnyak, Y. Koshevnik and E. A. Burstein, Decomposition of protein tryptophan fluorescence spectra into log-normal components. III. Correlation between fluorescence and microenvironment parameters of individual tryptophan residues, *Biophys. J.*, 2001, **81**, 1735–1758.
[108] J. Gsponer, U. Haberthur and A. Caflisch, The role of side-chain interactions in the early steps of aggregation: Molecular dynamics simulations of an amyloid-forming peptide from the yeast prion Sup35, *Proc. Natl. Acad. Sci.*, 2003, **100**, 5154–5159.
[109] G. Colombo, D. Roccatano and A. E. Mark, Folding and stability of the three-stranded β-sheet peptide Betanova: Insights from molecular dynamics simulations, *Proteins*, 2002, **46**, 380–392.
[110] R. B. Best, B. Li, A. Steward, V. Daggett and J. Clarke, Can non-mechanical proteins withstand force? Stretching barnase by atomic force microscopy and molecular dynamics simulation, *Biophys. J.*, 2001, **81**, 2344–2356.
[111] Y. Fan, L. M. Shi, K. W. Kohn, Y. Pommier and J. N. Weinstein, Quantitative structure-antitumor activity relationships of camptothecin analogues: Cluster analysis and genetic algorithm-based studies, *J. Med. Chem.*, 2001, **44**, 3254–3263.
[112] F. A. Hamprecht, C. Peter, X. Daura, W. Thiel and W. F. van Gunsteren, A strategy for analysis of (molecular) equilibrium simulations: Configuration space density estimation, clustering, and visualization, *J. Chem. Phys.*, 2001, **114**, 2079–2089.
[113] M. Vankatarajan and W. Braun, New quantitative descriptors of amino acids based on multidimensional scaling of a large number of physical chemical properties, *J. Mol. Mod.*, 2001, **7**, 445–453.
[114] Z. Zhang, Y. Zhu and Y. Shi, Molecular dynamics simulations of urea and thermal-induced denaturation of S-peptide analogue, *Biophys. Chem.*, 2001, **89**, 145–162.
[115] L. Carlacci, Conformational analysis of a farnesyltransferase peptide inhibitor, *CVIM, J. Comput.-Aided Mol. Des.*, 2000, **14**, 369–382.
[116] P. Ferrara, J. Apostolakis and A. Caflisch, Thermodynamics and kinetics of folding of two model peptides investigated by molecular dynamics simulations, *J. Phys. Chem. B*, 2000, **104**, 5000–5010.
[117] D. K. Klimov and D. Thirumalai, *M*echanisms and kinetics of β-hairpin formation, *Proc. Natl. Acad. Sci.*, 2000, **97**, 2544–2549.

[118] A. Li and V. Daggett, Identification and characterization of the unfolding transition state of chymotrypsin inhibitor 2 by molecular dynamics simulations, *J. Mol. Biol.*, 1996, **247**, 412–419.
[119] S. Mariappan, A. Garcoa and G. Gupta, Structure and dynamics of the DNA hairpins formed by tandemly repeated CTG triplets associated with myotonic dystrophy, *Nucleic Acids Res.*, 1996, **24**, 775–783.
[120] E. M. Boczko and C. L. Brooks III, First-principle calculation of the folding free energy of a three-helix bundle protein, *Science*, 1995, **269**, 393–396.
[121] A. Li and V. Daggett, Characterization of the transition state of protein unfolding by use of molecular dynamics: Chymotrypsin inhibitor 2, *Proc. Natl. Acad. Sci.*, 1994, **91**, 10430–10434.
[122] M. E. Karpen, D. J. Tobias and C. L. Brooks III, Statistical clustering techniques for the analysis of long molecular dynamics trajectories: Analysis of 2.2 ns trajectories of YPGDV, *Biochemistry*, 1993, **32**, 412–420.
[123] N. Bruant, D. Flatters, R. Lavery and D. Genest, From atomic to mesoscopic descriptions of the internal dynamics of DNA, *Biophys. J.*, 1999, **77**, 2366–2376.
[124] D. Genest, Correlated motions analysis from molecular dynamics trajectories: Statistical accuracy on the determination of canonical correlation coefficients, *J. Comp. Chem.*, 1999, **20**, 1571–1576.
[125] D. Genest, Motion of groups of atoms in DNA studied by molecular dynamics simulation, *Eur. Biophys. J.*, 1998, **27**, 283–289.
[126] Y. Xia and M. Levitt, Funnel-like organization in sequence space determines the distributions of protein stability and folding rate preferred by evolution, *Proteins*, 2004, **55**, 107–114.
[127] O. Ivanciuc, C. H. Schein and W. Braun, SDAP: Database and computational tools for allergenic proteins, *Nucleic Acids Res.*, 2003, **31**, 359–362.
[128] D. Mihailescu, J. Reed and J. C. Smith, Convergence in peptide folding simulation: Multiple trajectories of a potential AIDS pharmacophore, *Biopolymers*, 2003, **70**, 121–133.
[129] G. E. Sims and S.-H. Kim, Global mapping of nucleic acid conformational space: Dinucleoside monophosphate conformations and transition pathways among conformational classes, *Nucleic Acids Res.*, 2003, **31**, 5607–5616.
[130] M. Feher and J. M. Schmidt, Metric and multidimensional scaling: Efficient tools for clustering molecular conformations, *J. Chem. Inf. Comput. Sci.*, 2001, **41**, 346–353.
[131] K. Pearson, On lines and planes of closest fit to a system of points in space, *Phil. Mag.*, 1901, **2**, 559–572.
[132] H. Hotelling, Analysis of a complex of statistical variables into principal components, *J. Educ. Psych.*, 1933, **24**, 417–441.
[133] J. Roach, S. Sharma, M. Kapustina and C. W. Carter, Structure alignment via Delauney tetrahedralization, *Proteins*, 2005, **60**, 66–81.
[134] V. Alexandrov and M. Gerstein, Using 3D hidden Markov models that explicitly represent spatial coordinates to model and compare protein structures, *BMC Bioinform.*, 2004, 5(2).
[135] T. R. Scheider, Domain identification by iterative analysis of error-scaled difference distance matrices, *Acta Cryst. D*, 2004, **60**, 2269–2275.
[136] Y. Ye and A. Godzik, Database searching by flexible protein structure alignment, *Protein Sci.*, 2004, **13**, 1841–1850.
[137] A. I. Jewett, C. C. Huang and T. E. Ferrin, MINRMS: An efficient algorithm for determining protein structure similarity using root-mean-squared-distance, *Bioinformatics*, 2003, **19**, 625–634.
[138] V. Kotlovyi, W. L. Nichols and L. F. T. Eyck, Protein structural alignment for detection of maximally conserved regions, *Biophys. Chem.*, 2003, **105**, 595–608.
[139] T. R. Scheider, A genetic algorithm for the identification of conformationally invariant regions in protein molecules, *Acta Cryst. D*, 2002, **58**, 195–208.

[140] M. Shatsky, R. Nussinov and H. J. Wolfson, Flexible protein alignment and hinge detection, *Proteins*, 2002, **48**, 242–256.
[141] J. A. Irving, J. C. Whisstock and A. M. Lesk, Protein structural alignments and functional genomics, *Proteins*, 2001, **42**, 378–382.
[142] J. A. Cuff and G. J. Barton, Application of multiple sequence alignment profiles to improve protein secondary structure prediction, *Proteins*, 2000, **40**, 502–511.
[143] W. G. Krebs and M. Gerstein, The morph server: a standardized system for analyzing and visualizing macromolecular motions in a database framework, *Nucleic Acids Res.*, 2000, **28**, 1665–1675.
[144] C. Notredame, D. G. Higgins and J. Heringa, T-coffee: A novel method for fast and accurate multiple sequence alignment, *J. Mol. Biol.*, 2000, **302**, 205–217.
[145] A. F. Neuwald, J. S. Liu, D. J. Lipman and C. E. Lawrence, Extracting protein alignment models from the sequence database, *Nucl. Acids Res.*, 1997, **25**, 1665–1677.
[146] W. L. Nichols, B. H. Zimm and L. F. T. Eyck, Conformation-invariant structures of the a1b1 human hemoglobin dimmer, *J. Mol. Biol.*, 1997, **270**, 598–615.
[147] W. Wriggers and K. Schulten, Protein domain movements: Detection of rigid domains and visualization of hinges in comparisons of atomic coordinates, *Proteins*, 1997, **29**, 1–14.
[148] M. Gerstein and R. B. Altman, Average core structures and variability measures for protein families: Application to the immunoglobins, *J. Mol. Biol.*, 1995, **251**, 161–175.
[149] J. Hein, An algorithm combining DNA and protein alignment, *J. Theor. Biol.*, 1994, **167**, 169–174.
[150] H. Carlson, Personal Communication, 2005.
[151] J. Johnson, D. Lamb, H. Frauenfelder, J. Muller, B. McMahon, G. Nienhaus and R. Young, Ligand binding to heme proteins. VI. Interconversion of taxonomic substates in carbonmonoxymyoglobin, *Biophys. J.*, 1996, **71**, 1563–1573.
[152] W. D. Tian, J. T. Sage, P. M. Champion, E. Chien and S. G. Sligar, Probing heme protein conformational equilibration rates with kinetic selection, *Biochemistry*, 1996, **35**, 3487–3502.
[153] T. Li, M. L. Quillin, G. N. Phillips, Jr. and J. S. Olson, Structural determinants of the stretching frequency of CO bound to myoglobin, *Biochemistry*, 1994, **33**, 1433–1446.
[154] S. Balasubramanian, D. G. Lambright, M. C. Marden and S. G. Boxer, Carbon monoxide recombination to human myoglobin mutants in glycerol-water solutions, *Biochemistry*, 1993, **32**, 2202–2212.
[155] D. Braunstein, K. Chu, K. Egeberg, H. Frauenfelder, J. Mourant, G. Nienhaus, P. Ormos, S. Sligar, B. Springer and R. Young, Ligand binding to heme proteins: III. FTIR studies of His-E7 and Val-E11 mutants of carbonmonoxymyoglobin, *Biophys. J.*, 1993, **65**, 2447–2454.
[156] K. M. Elsawy, M. K. Hodgson and L. S. D. Caves, The physical determinants of the DNA conformational landscape: an analysis of the potential energy surface of single-strand dinucleotides in the conformational space of duplex DNA, *Nucleic Acids Res*, 2005, **33**, 5749–5762.
[157] R. Cattell, The meaning and strategic use of factor analysis. In *Handbook of Multivariate Experimental Psychology* (ed. R. B. Cattell), Rand McNally, Chicago, 1966, pp. 174–243.
[158] R. B. Cattell, The scree test for the number of factors, *Multivar. Behav. Res.*, 1966, **1**, 245–276.
[159] D. J. Wales, *Energy landscapes with applications to clusters, biomolecules and glasses*, Cambridge University Press, Cambridge, 2003.
[160] R. Kazmierkiewicz, C. Czaplewski, B. Lammek and J. Ciarkowski, Essential dynamics/factor analysis for the interpretation of molecular dynamics trajectories, *J. Comput.-Aided Mol. Des.*, 1999, **13**, 21–33.

[161] A. Amadei, A. B. Linssen and H. J. Berendsen, Essential dynamics of proteins, *Proteins*, 1993, **17**, 412–425.
[162] A. Noy, T. Meyer, M. Rueda, C. Ferrer, A. Valencia, A. Perez, X. d. l. Cruz, J. M. Lopez-Bes, R. Pouplana, J. Fernandez-Recio, F. J. Luque and M. Orozco, Data mining of molecular dynamics trajectories of nucleic acids, *J. Biomol. Struct. Dyn.*, 2006, **23**, 447–455.
[163] S. Nunez, C. Wing, D. Antoniou, V. L. Schramm and S. D. Schwartz, Insight into catalytically relevant correlated motions in human purine nucleoside phosphorylase, *J. Phys. Chem. A*, 2006, **110**, 463–472.
[164] N. J. Marianayagam and S. E. Jackson, Native-state dynamics of the ubiquitin family: Implications for function and evolution, *J. Royal Soc. Interface*, 2005, **2**, 47–54.
[165] A. Perez, J. R. Blas, M. Rueda, J. M. Lopez-Bes, X. De la Cruz and M. Orozco, Exploring the essential dynamics of B-DNA, *J. Chem. Theory Comput.*, 2005, **1**, 790–800.
[166] K. Arora and T. Schlick, In silico evidence for DNA polymerase-beta's substrate-induced conformational change, *Biophys. J.*, 2004, **87**, 3088–3099.
[167] J. E. Ollerenshaw, H. Kaya, H. S. Chan and L. E. Kay, Sparsely populated folding intermediates of the Fyn SH3 domain: Matching native-centric essential dynamics and experiment, *Proc. Natl. Acad. Sci.*, 2004, **101**, 14748–14753.
[168] L. V. Mello, B. L. De Groot, S. Li and M. J. Jedrzejas, Structure and flexibility of Streptococcus agalactiae hyaluronate lyase complex with its substrate. Insights into the mechanism of processive degradation of hyaluronan, *J. Biol. Chem.*, 2002, **277**, 36678–36688.
[169] R. Yang, M. C. Lee, H. Yan and Y. Duan, Loop conformation and dynamics of the Escherichia coli HPPK apo-enzyme and its binary complex with MgATP, *Biophys. J.*, 2005, **89**, 95–106.
[170] D. Komander, S. Kular Gursant, W. Schuttelkopf Alexander, M. Deak, K. R. C. Prakash, J. Bain, M. Elliott, M. Garrido-Franco, P. Kozikowski Alan, R. Alessi Dario and M. F. van Aalten Daan, Interactions of LY333531 and other bisindolyl maleimide inhibitors with PDK1, *Structure*, 2004, **12**, 215–226.
[171] P. Barthe, C. Roumestand, H. Demene and L. Chiche, Helix motion in protein C12A-p8MTCP1: Comparison of molecular dynamics simulations and multifield NMR relaxation data, *J. Comp. Chem.*, 2002, **23**, 1577–1586.
[172] R. M. Biondi, D. Komander, C. C. Thomas, J. M. Lizcano, M. Deak, D. R. Alessi and D. M. F. van Aalten, High resolution crystal structure of the human PDK1 catalytic domain defines the regulatory phosphopeptide docking site, *EMBO J.*, 2002, **21**, 4219–4228.
[173] D. M. F. van Aalten, C. R. Chong and L. Joshua-Tor, Crystal structure of carboxypeptidase A complexed with D-cysteine at 1.75-Å–inhibitor-induced conformational changes,, *Biochemistry*, 2000, **39**, 10082–10089.
[174] D. M. F. van Aalten, W. Crielaard, K. J. Hellingwerf and L. Joshua-Tor, Conformational substates in different crystal forms of the photoactive yellow protein – Correlation with theoretical and experimental flexibility, *Protein Sci.*, 2000, **9**, 64–72.
[175] S. Hayward, Structural principles governing domain motions in proteins, *Proteins*, 1999, **36**, 425–435.
[176] R. Abseher, L. Horstink, C. W. Hilbers and M. Nilges, Essential spaces defined by NMR structure ensembles and molecular dynamics simulation show significant overlap, *Proteins*, 1998, **31**, 370–382.
[177] B. L. de Groot, S. Hayward, D. M. van Aalten, A. Amadei and H. J. Berendsen, Domain motions in bacteriophage T4 lysozyme: a comparison between molecular dynamics and crystallographic data, *Proteins*, 1998, **31**, 116–127.
[178] L. Ragona, G. Colombo, M. Catalano and H. Molinari, Determinants of protein stability and folding: Comparative analysis of β-lactoglobulins and liver basic fatty acid binding protein, *Proteins*, 2005, **61**, 366–376.

[179] Y. Sugita and Y. Okamoto, Molecular mechanism for stabilizing a short helical peptide studied by generalized-ensemble simulations with explicit solvent, *Biophys. J.*, 2005, **88**, 3180–3190.
[180] A. Merlino, G. Graziano and L. Mazzarella, Structural and dynamic effects of α-helix deletion in Sso7d: Implications for protein thermal stability, *Proteins*, 2004, **57**, 692–701.
[181] D. Roccatano, I. Daidone, M.-A. Ceruso, C. Bossa and D. Nola Alfredo, Selective excitation of native fluctuations during thermal unfolding simulations: Horse heart cytochrome c as a case study, *Biophys. J.*, 2003, **84**, 1876–1883.
[182] J. Lee and S. Shin, Two-dimensional correlation analysis of peptide unfolding: Molecular dynamics simulations of β hairpins, *J. Phys. Chem. B*, 2002, **106**, 8796–8802.
[183] B. L. de Groot, X. Daura, A. E. Mark and H. Grubmuller, Essential dynamics of reversible peptide folding: Memory-free conformational dynamics governed by internal hydrogen bonds, *J. Mol. Biol.*, 2001, **309**, 299–313.
[184] L. D. Creveld, A. Amadei, R. C. van Schaik, H. A. Pepermans, J. de Vlieg and H. J. Berendsen, Identification of functional and unfolding motions of cutinase as obtained from molecular dynamics computer simulations, *Proteins*, 1998, **33**, 253–264.
[185] R. J. Law, R. H. Henchman and J. A. McCammon, A gating mechanism proposed from a simulation of a human α7 nicotinic acetylcholine receptor, *Proc. Natl. Acad. Sci.*, 2005, **102**, 6813–6818.
[186] A. Grottesi and S. P. Sansom Mark, Molecular dynamics simulations of a K+ channel blocker: Tc1 toxin from Tityus cambridgei, *FEBS lett*, 2003, **535**, 29–33.
[187] D. P. Tieleman, B. Hess and M. S. P. Sansom, Analysis and evaluation of channel models: Simulations of alamethicin, *Biophys. J.*, 2002, **83**, 2393–2407.
[188] R. D. Lins and T. P. Straatsma, Computer simulation of the rough lipopolysaccharide membrane of Pseudomonas aeruginosa, *Biophys. J.*, 2001, **81**, 1037–1046.
[189] G. H. Peters and R. P. Bywater, Influence of a lipid interface on protein dynamics in a fungal lipase, *Biophys. J.*, 2001, **81**, 3052–3065.
[190] I. H. Shrivastava, C. E. Capener, L. R. Forrest and M. S. P. Sansom, Structure and dynamics of K channel pore-lining helices: A comparative simulation study, *Biophys. J.*, 2000, **78**, 79–92.
[191] D. Cregut, G. Drin, J. P. Liautard and L. Chiche, Hinge-bending motions in annexins: Molecular dynamics and essential dynamics of apo-annexin V and of calcium bound annexin V and I, *Protein Eng.*, 1998, **11**, 891–900.
[192] M. C. Lee, J. Deng, J. M. Briggs and Y. Duan, Large-scale conformational dynamics of the HIV-1 integrase core domain and its catalytic loop mutants, *Biophys. J.*, 2005, **88**, 3133–3146.
[193] I. Daidone, D. Roccatano and S. Hayward, Investigating the accessibility of the closed domain conformation of citrate synthase using essential dynamics sampling, *J. Mol. Biol.*, 2004, **339**, 515–525.
[194] I. Stoica, Solvent interactions and protein dynamics in spin-labeled T4 lysozyme, *J. Biomol. Struct. Dyn.*, 2004, **21**, 745–760.
[195] N. E. Labrou, L. V. Mello and Y. D. Clonis, Functional and structural roles of the glutathione-binding residues in maize (Zea mays) glutathione S-transferase I, *Biochem. J.*, 2001, **358**, 101–110.
[196] D. Roccatano, A. E. Mark and S. Hayward, Investigation of the mechanism of domain closure in citrate synthase by molecular dynamics simulation, *J. Mol. Biol.*, 2001, **310**, 1039–1053.
[197] T. A. Soares, J. H. Miller and T. P. Straatsma, Revisiting the structural flexibility of the complex p21ras-GTP: The catalytic conformation of the molecular switch II, *Proteins*, 2001, **45**, 297–312.
[198] C. R. Watts, G. Toth, R. F. Murphy and S. Lovas, Domain movement in the epidermal growth factor family of peptides, *Theochemistry*, 2001, **535**, 171–182.

[199] D. M. F. van Aalten, P. C. Jones, M. De Sousa and J. B. C. Findlay, Engineering protein mechanics: Inhibition of concerted motions of the cellular retinol binding protein by site-directed mutagenesis, *Protein Eng.*, 1997, **10**, 31–37.
[200] C. Arcangeli, A. R. Bizzarri and S. Cannistraro, Molecular dynamics simulation and essential dynamics study of mutated plastocyanin: Structural, dynamical and functional effects of a disulfide bridge insertion at the protein surface, *Biophys. Chem.*, 2001, **92**, 183–199.
[201] L. Stella, E. E. Di Iorio, M. Nicotra and G. Ricci, Molecular dynamics simulations of human glutathione transferase P1-1: Conformational fluctuations of the apo-structure, *Proteins*, 1999, **37**, 10–19.
[202] A. Pandini and L. Bonati, Conservation and specialization in PAS domain dynamics, *Prot. Eng. Des. Sel.*, 2005, **18**, 127–137.
[203] A. Merlino, L. Vitagliano, A. Ceruso Marc and L. Mazzarella, Subtle functional collective motions in pancreatic-like ribonucleases: From ribonuclease A to angiogenin, *Proteins*, 2003, **53**, 101–110.
[204] M. Sulpizi, U. Rothlisberger and P. Carloni, Molecular dynamics studies of caspase-3, *Biophys. J.*, 2003, **84**, 2207–2215.
[205] A. Grottesi, M.-A. Ceruso, A. Colosimo and A. Di Nola, Molecular dynamics study of a hyperthermophilic and a mesophilic rubredoxin, *Proteins*, 2002, **46**, 287–294.
[206] M. J. Jedrzejas, L. V. Mello, B. de Groot and S. Li, Mechanism of hyaluronan degradation by *Streptococcus pneumoniae* hyaluronate lyase, *J. Biol. Chem.*, 2002, **277**, 28287–28297.
[207] M. Otyepka and J. Damborsky, Functionally relevant motions of haloalkane dehalogenases occur in the specificity-modulating cap domain, *Protein Sci.*, 2002, **11**, 1206–1217.
[208] B. S. Sanjeev and S. Vishveshwara, Essential dynamics and sidechain hydrogen bond cluster studies on eosinophil cationic protein, *Eur. Phys. J. D.*, 2002, **20**, 601–608.
[209] C. Arcangeli, A. R. Bizzarri and S. Cannistraro, Concerted motions in copper plastocyanin and azurin: An essential dynamics study, *Biophys. Chem.*, 2001, **90**, 45–56.
[210] R. D. Lins, T. P. Straatsma and J. M. Briggs, Similarities in the HIV-1 and ASV integrase active sites upon metal cofactor binding, *Biopolymers*, 2000, **53**, 308–315.
[211] P. L. Chau, D. M. F. van Aalten, R. P. Bywater and J. B. C. Findlay, Functional concerted motions in the bovine serum retinol-binding protein, *J. Comput.-Aided Mol. Des.*, 1999, **13**, 11–20.
[212] B. L. de Groot, G. Vriend and H. J. Berendsen, Conformational changes in the chaperonin GroEL: New insights into the allosteric mechanism, *J. Mol. Biol.*, 1999, **286**, 1241–1249.
[213] L. Horstink, R. Abseher, M. Nilges and C. W. Hilbers, Functionally important correlated motions in the single-stranded DNA-binding protein encoded by filamentous phage Pf3, *J. Mol. Biol.*, 1999, **287**, 569–577.
[214] R. D. Lins, J. M. Briggs, T. P. Straatsma, H. A. Carlson, J. Greenwald, S. Choe and J. A. McCammon, Molecular dynamics studies on the HIV-1 integrase catalytic domain, *Biophys. J.*, 1999, **76**, 2999–3011.
[215] D. M. F. van Aalten, W. D. Hoff, J. B. C. Findlay, W. Crielaard and K. J. Hellingwerf, Concerted motions in the photoactive yellow protein, *Protein Eng.*, 1998, **11**, 873–879.
[216] A. Merlino, L. Vitagliano, M. A. Ceruso and L. Mazzarella, Dynamic properties of the N-terminal swapped dimer of ribonuclease A, *Biophys. J.*, 2004, **86**, 2383–2391.
[217] M. A. Ceruso, A. Grottesi and A. Di Nola, Dynamic effects of mutations within two loops of cytochrome c551 from Pseudomonas aeruginosa, *Proteins*, 2003, **50**, 222–229.

[218] G. Settanni, A. Cattaneo and P. Carloni, Molecular dynamics simulations of the NGF-TrkA domain 5 complex and comparison with biological data, *Biophys. J.*, 2003, **84**, 2282–2292.
[219] J. Lee, K. Lee and S. Shin, Theoretical studies of the response of a protein structure to cavity-creating mutations, *Biophys. J.*, 2000, **78**, 1665–1671.
[220] A. Brigo, K. W. Lee, G. I. Mustata and J. M. Briggs, Comparison of multiple molecular dynamics trajectories calculated for the drug-resistant HIV-1 integrase T66I/M154I catalytic domain, *Biophys. J.*, 2005, **88**, 3072–3082.
[221] K. S. Chakrabarti, B. S. Sanjeev and S. Vishveshwara, Stability and dynamics of domain-swapped bovine-seminal ribonuclease, *Chem. Biodiv.*, 2004, **1**, 802–818.
[222] R. G. Efremov, Y. A. Kosinsky, D. E. Nolde, R. Tsivkovskii, A. S. Arseniev and S. Lutsenko, Molecular modelling of the nucleotide-binding domain of Wilson's disease protein: Location of the ATP-binding site, domain dynamics, and potential effects of the major disease mutations, *Biochem. J.*, 2004, **382**, 293–305.
[223] F. Fraternali, L. Cavallo and G. Musco, Effects of pathological mutations on the stability of a conserved amino acid triad in retinoschisin, *FEBS Lett.*, 2003, **544**, 21–26.
[224] M. L. Sforca, S. Oyama, Jr., F. Canduri, C. C. B. Lorenzi, T. A. Pertinhez, K. Konno, B. M. Souza, M. S. Palma, N. J. Ruggiero, W. F. Azevedo, Jr. and A. Spisni, How C-terminal carboxyamidation alters the biological activity of peptides from the venom of the eumenine solitary wasp, *Biochemistry*, 2004, **43**, 5608–5617.
[225] A. Crespo, A. Marti Marcelo, G. Kalko Susana, A. Morreale, M. Orozco, L. Gelpi Jose, F. J. Luque and A. Estrin Dario, Theoretical study of the truncated hemoglobin HbN: Exploring the molecular basis of the NO detoxification mechanism, *J. Am. Chem. Soc.*, 2005, **127**, 4433–4444.
[226] M. Fuxreiter, C. Magyar, T. Juhasz, Z. Szeltner, L. Polgar and I. Simon, Flexibility of prolyl oligopeptidase: Molecular dynamics and molecular framework analysis of the potential substrate pathways, *Proteins*, 2005, **60**, 504–512.
[227] C. Bossa, M. Anselmi, D. Roccatano, A. Amadei, B. Vallone, M. Brunori and A. Di Nola, Extended molecular dynamics simulation of the carbon monoxide migration in sperm whale myoglobin, *Biophys. J.*, 2004, **86**, 3855–3862.
[228] M. L. Barreca, K. W. Lee, A. Chimirri and J. M. Briggs, Molecular dynamics studies of the wild-type and double mutant HIV-1 integrase complexed with the 5CITEP inhibitor: Mechanism for inhibition and drug resistance, *Biophys. J.*, 2003, **84**, 1450–1463.
[229] G. I. Mustata, T. A. Soares and J. M. Briggs, Molecular dynamics studies of alanine racemase: A structural model for drug design, *Biopolymers*, 2003, **70**, 186–200.
[230] A. Pang, Y. Arinaminpathy, S. P. Sansom Mark and C. Biggin Philip, Interdomain dynamics and ligand binding: Molecular dynamics simulations of glutamine binding protein, *FEBS lett.*, 2003, **550**, 168–174.
[231] S. K. Ludemann, V. Lounnas and R. C. Wade, How do substrates enter and products exit the buried active site of cytochrome P450cam? 2. Steered molecular dynamics and adiabatic mapping of substrate pathways, *J. Mol. Biol.*, 2000, **303**, 813–830.
[232] G. H. Peters, T. M. Frimurer, J. N. Andersen and O. H. Olsen, Molecular dynamics simulations of protein-tyrosine phosphatase 1B. II. Substrate-enzyme interactions and dynamics, *Biophys. J.*, 2000, **78**, 2191–2200.
[233] G. H. Peters, D. M. van Aalten, A. Svendsen and R. Bywater, Essential dynamics of lipase binding sites: The effect of inhibitors of different chain length, *Protein Eng.*, 1997, **10**, 149–158.
[234] I. Eberini, A. M. Baptista, E. Gianazza, F. Fraternali and T. Beringhelli, Reorganization in apo- and holo-β-lactoglobulin upon protonation of Glu89: Molecular dynamics and pKa calculations, *Proteins*, 2004, **54**, 744–758.

[235] R. Gargallo, P. H. Huenenberger, F. X. Aviles and B. Oliva, Molecular dynamics simulation of highly charged proteins: Comparison of the particle-particle particle-mesh and reaction field methods for the calculation of electrostatic interactions, *Protein Sci.*, 2003, **12**, 2161–2172.
[236] D. Mustard and W. Ritchie David, Docking essential dynamics eigenstructures, *Proteins*, 2005, **60**, 269–274.
[237] S. Sharma, P. Pirila, H. Kaija, K. Porvari, P. Vihko and A. H. Juffer, Theoretical investigations of prostatic acid phosphatase, *Proteins*, 2005, **58**, 295–308.
[238] S. Ferrari, P. M. Costi and R. C. Wade, Inhibitor specificity via protein dynamics insights from the design of antibacterial agents targeted against thymidylate synthase, *Chem. Biol.*, 2003, **10**, 1183–1193.
[239] A. Noy, A. Perez, F. Lankas, F. J. Luque and M. Orozco, Relative flexibility of DNA and RNA: A molecular dynamcis study, *J. Mol. Biol.*, 2004, **343**, 627–638.
[240] A. Perez, A. Noy, F. Lankas, F. J. Luque and M. Orozco, The relative flexibility of B-DNA and A-RNA duplexes: Database analysis, *Nucleic Acids Res.*, 2004, **32**, 6144–6151.
[241] M. Orozco, A. Perez, A. Noy and F. J. Luque, Theoretical methods for the simulation of nucleic acids, *Chem. Soc. Rev.*, 2003, **32**, 350–364.
[242] V. Cojocaru, R. Klement and T. M. Jovin, Loss of G-A base pairs is insufficient for achieving a large opening of U4 snRNA K-turn motif, *Nucleic Acids Res.*, 2005, **33**, 3435–3446.
[243] A. Noy, A. Perez, M. Marquez, F. J. Luque and M. Orozco, Structure, recognition properties, and flexibility of the DNA.RNA hybrid, *J. Am. Chem. Soc.*, 2005, **127**, 4910–4920.
[244] R. Soliva, V. Monaco, I. Gomez-Pinto, N. J. Meeuwenoord, G. A. V. d. Marel, J. H. V. Boom, C. Gonzalez and M. Orozco, Solution structure of a DNA duplex with a chiral alkyl phosphonate moiety, *Nucleic Acids Res.*, 2001, **29**, 2973–2985.
[245] A. Ninaber and J. M. Goodfellow, DNA conformation and dynamics, *Radiat. Environ. Biophys.*, 1999, **38**, 23–29.
[246] H. Yamaguchi, D. M. F. Van Aalten, M. Pinak, A. Furukawa and R. Osman, Essential dynamics of DNA containing a cis.syn cyclobutane thymine dimer lesion, *Nucleic Acids Res.*, 1998, **26**, 1939–1946.
[247] H. Yamaguchi, J. G. Siebers, A. Furukawa, N. Otagiri and R. Osman, Molecular dynamics simulation of a DNA containing a single strand break, *Radiat. Prot. Dosimetry.*, 2002, **99**, 103–108.
[248] D. M. F. van Aalten, D. A. Erlanson, G. L. Verdine and L. Joshua-Tor, A structural snapshot of base-pair opening in DNA, *Proc. Natl. Acad. Sci.*, 1999, **96**, 11809–11814.
[249] S. Jha, P. V. Coveney and C. A. Laughton, Force field validation for nucleic acid simulations: Comparing energies and dynamics of a DNA dodecamer, *J. Comp. Chem.*, 2005, **26**, 1617–1627.
[250] M. Rueda, S. G. Kalko, F. J. Luque and M. Orozco, The structure and dynamics of DNA in the gas phase, *J. Am. Chem. Soc.*, 2003, **125**, 8007–8014.
[251] V. Tsui and D. A. Case, Molecular dynamics simulations of nucleic acids with a generalized born model, *J. Am. Chem. Soc.*, 2000, **122**, 2489–2498.
[252] M. G. Gustafsson, Independent component analysis yields chemically interpretable latent variables in multivariate regression, *J. Chem. Inf. Model.*, 2005, **45**, 1244–1255.
[253] P. Comon, Independent component analysis, *a new concept?, Signal Processing*, 1994, **36**, 287–314.
[254] A. Hyvarinen, Survey on independent component analysis, *Neural Comput. Surveys*, 1999, **2**, 94–128.
[255] R. D. S. Yadava and R. Chaudhary, Solvation Transduction and independent component analysis for pattern recognition in SAW electronic nose, *Sens. Actuators B*, 2006, **113**, 1–21.

[256] F. Westad and M. Kermit, Cross validation and uncertainty estimates in independent component analysis, *Anal. Chim. Acta.*, 2003, **490**, 341–354.
[257] F. S. de Edelenyi, A. W. Simonetti, G. Postma, R. Huo and L. M. C. Buydens, Application of independent component analysis to 1H MR spectroscopic imaging exams of brain tumors, *Anal. Chim. Acta.*, 2005, **544**, 36–46.
[258] A. Pichler and M. G. Sowa, Blind phase projection as an effective means of recovering pure component spectra from phase modulated photoacoustic spectra, *Vib. Spectrosc.*, 2005, **39**, 163–168.
[259] M. Alrubaiee, M. Xu, S. K. Gayen, M. Brito and R. R. Alfano, Three-dimensional optical tomographic imaging of scattering objects in tissue-simulating turbid media using independent component analysis, *Appl. Phys. Lett.*, 2005, **87**, 191112–191113.
[260] E. De Lauro, S. De Martino, M. Falanga, A. Ciaramella and R. Tagliaferri, Complexity of time series associated to dynamical systems inferred from independent component analysis, *Phys. Rev. E*, 2005, **72**, 046712/1–046712/14.
[261] S. De Martino, M. Falanga and L. Mona, Stochastic resonance mechanism in aerosol index dynamics, *Phys. Rev. Lett.*, 2002, **89**, 128501/1–128501/4.
[262] G. W. Stewart, On the early history of the singular value decomposition, *SIAM Rev*, 1993, **35**, 551–566.
[263] M. E. Wall, A. Rechtsteiner and L. M. Rocha, Singular value decomposition and principal component analysis. In *A Practical Approach to Microarray Data Analysis* (eds D. P. Berrar, W. Dubitzky and M. Granzow), Kluwer, Norwell, MA, 2003, pp. 91–109.
[264] J. Tomfohr, J. Lu and T. B. Kepler, Pathway level analysis of gene expression using singular value decomposition, *BMC Bioinform.* 6 (2005).
[265] J. Walton and N. Fairley, Noise reduction in X-ray photoelectron spectromicroscopy by a singular value decomposition sorting procedure, *J. Electr. Spectr. Relat. Phenom.*, 2005, **148**, 29–40.
[266] N. Trbovic, S. Smirnov, F. Zhang and R. Bruschweiler, Covariance NMR spectroscopy by singular value decomposition, *J. Magn. Reson.*, 2004, **171**, 277–283.
[267] S. Maguid, S. Fernandez-Alberti, L. Ferrelli and J. Echave, Exploring the common dynamics of homologous proteins. Application to the globin family, *Biophys. J.*, 2005, **89**, 3–13.
[268] J. A. Hanley, Appropriate uses of multivariate analysis, *Annu. Rev. Public Health*, 1983, **4**, 155–180.