WILEY InterScience® DISCOVER SOMETHING GREAT

PROTEINS STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

Full correlation analysis of conformational protein dynamics

Oliver F. Lange and Helmut Grubmüller*

Department of Theoretical and Computational Biophysics, Max-Planck-Institute for Biophysical Chemistry, Am Fassberg 11, Göttingen 37077, Germany

ABSTRACT

Correlated motions in biomolecules are often essential for their function, for example, allosteric signal transduction or mechanical/thermodynamic energy transport. Principal component analysis (PCA) is a widely used method to extract functionally relevant collective motions from a molecular dynamics (MD) trajectory. Being based on the covariance matrix, however, PCA detects only linear correlations. Here we present a new method, full correlation analysis (FCA), which is based on mutual information and thus quantifies all correlations, including nonlinear and higher order correlations. For comparison, we applied both, PCA and FCA, to ~100 ns MD trajectories of T4 lysozyme and the hexapeptide neurotensin. For both systems, FCA yielded better resolved conformational substates and aligned its modes more often with actual transition pathways. This improved resolution is shown to be due to a strongly increased anharmonicity of FCA modes as compared to the respective PCA modes. The high anharmonicity further suggests that the motions extracted by FCA are functionally more relevant than those captured by PCA. In summary, FCA should provide improved collective degrees of freedom for dimension-reduced descriptions of macromolecular dynamics.

Proteins 2008; 70:1294–1312.

Key words: MD simulation; independent component analysis; principal component analysis; correlated motion; collective motion; protein dynamics; conformational dynamics; lysozyme; mutual information.

INTRODUCTION

Collective motions of proteins are essential in many respects for protein function, such as substrate binding and product release, regulation and allosteric behavior, as well as contractile and motor functions.[1] Although recent computational and algorithmic advances enable us to simulate protein dynamics accurately in realistic environments over hundreds of nanoseconds,[2] extracting the functionally relevant collective motions from the molecular dynamics (MD) trajectory still poses a considerable challenge.[3] The two most widely used methods to determine these motions are normal mode analysis (NMA)[4–6] and principal component analysis (PCA).[7–9]

These two methods differ mainly in their respective assumptions what constitutes functionally relevant motions. Since many functional processes involve large and slow conformational changes (as opposed to small-amplitude fast thermal vibrations), PCA selects those collective degrees of freedom that contribute most to the total atomic displacements seen in the trajectory. NMA, in contrast, is motivated by the desire to obtain uncoupled degrees of freedom. Because of the required harmonic approximation, however, this separation is only very local in phase space, and anharmonic motions are not captured well. Accordingly, for small molecules, and given a sufficiently accurate force field or QM treatment, this approach reliably predicts infrared vibrational spectra from the Hessian matrix, that is, the second derivatives of the potential energy, and it has also been successfully applied to calculate high frequency vibrational spectra of proteins.[10] To what extent such a harmonic approximation to a single local minimum of the potential energy surface can characterize functional motions governed by the very complex multiminima energy landscape of proteins,[11] however, is far from clear.

PCA partially circumvents this limitation, as it rests on the covariance matrix of atomic displacements rather than on the Hessian matrix. Accordingly, PCA is based on a multivariate Gaussian approximation to the (canonical) configuration space density of the system, and the principal components may be reinterpreted as (uncoupled) normal modes in an approximate harmonic (quasi-harmonic) free energy surface.[12,13] In contrast to NMA, this approximation is nonlocal, and thanks to the statistical mechanics approach PCA also captures motions that result from visits to multiple minima, which for applications to macromolecules is a major advantage of PCA over NMA.

Grant sponsor: Volkswagen foundation; Grant numbers: I/80436, I/80585.
*Correspondence to: Helmut Grubmüller, Department of Theoretical and Computational Biophysics, Max-Planck-Institute for Biophysical Chemistry, Am Fassberg 11, Göttingen 37077, Germany.
E-mail: hgrubmu@gwdg.de
Received 5 December 2006; Revised 17 April 2007; Accepted 7 May 2007
Published online 17 September 2007 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.21618

Probably unexpected at first sight, those collective modes that accumulate the largest fluctuation amplitude, are those obtained from diagonalizing the covariance matrix of atomic displacements.[14] As a result, PCA identifies exactly those collective modes whose covariances vanish. However, because the covariance matrix describes only linear correlations, nonlinear correlations between the PCA modes can—and often do—persist, as already pointed out by Amadei *et al.*[9] Furthermore, higher order (i.e., multicoordinate) correlations between three or more modes are also not detected.

Here we present a new method, full correlation analysis (FCA), to obtain collective degrees of freedom which minimize all, linear, non-linear, as well as multicoordinate correlations. Avoiding harmonic and linear approximations altogether, we combine the advantages of PCA—a nonlocal statistical mechanics approach—and NMA, to yield uncoupled collective coordinates. To this aim, we quantify the correlations by the measure of mutual information (MI), which, derived from information theory, captures in fact all types of correlation.[15] For maximally uncoupled collective coordinates, we suggest to minimize the MI of the whole system by selecting from all possible rotations in configurational space the coordinate transformation with lowest MI.

To implement FCA we find help in the signal processing field. There, minimization of mutual information is used to extract independent sources from mixed signals, for example, blind source separation (BSS)[16] or independent component analysis (ICA).[17,18] These algorithms differ from each other in three main aspects. First, the estimation of MI can be either cumulant based, parametric (e.g., FastICA[19]), or nonparametric (e.g., MILCA[20]). Second, for the minimization of MI, diverse methods like stochastic descent, gradient descent, or a direct solution of the normal equations (e.g., FastICA) have been applied. Third, the resulting coordinates can be linear (e.g., MILCA) or nonlinear (e.g., MISEP[21]). Combining and selecting suitable features from these available algorithms, we developed an algorithm tailored towards the specific needs of FCA of biomolecular dynamics. We consider here, as a first step, generalized correlations between linear collective coordinates, though FCA can in principle also be applied to extract non-linear coordinates.

In the Methods Section, we summarize basic properties of MI and present the minimization algorithm. In the Results Section we first evaluate our algorithm for a test-system with known solution. Next, FCA is applied to a 117 ns MD trajectory of the T4 bacteriophage lysozyme and to a 100 ns trajectory of the hexapeptide neurotensin, and compared to PCA. Here, our main evaluation criterion is the ability to reveal and resolve conformational substates. For neurotensin, we additionally investigate to what extent PCA and FCA provide low-dimensional free energy surfaces that accurately describe conformational transitions and thus are suitable essential coordinates. Subsequently, we quantified the differences of amplitude, collectivity, and anharmonicity of FCA and PCA modes, as well as the remaining coupling between pairs of modes. Convergence of FCA modes is finally assessed by comparison to FCA modes extracted from a multidimensional random walk.

MATERIALS AND METHODS

Definition of mutual information

We briefly summarize the definition of mutual information, for more details see Refs. 22 and 23. We consider a statistical ensemble in the $3N$ dimensional configurational space of protein configurations $\mathbf{r}$ with mean structure $\langle\mathbf{r}\rangle$. The displacement vector $\mathbf{x} = \mathbf{r} - \langle\mathbf{r}\rangle$ consists of independent components, that is, the fluctuations in the coordinates $(x_1, x_2, \ldots, x_{3N})$ are uncorrelated, if and only if

$$p(\mathbf{x}) = \prod_{i=1}^{3N} p_i(x_i), \tag{1}$$

where $p(\mathbf{x})$ denotes the canonical ensemble density $p(\mathbf{x}) = Z^{-1} \exp[-\beta V(\mathbf{x} + \langle\mathbf{r}\rangle)]$, with partition function Z, inverse temperature β, potential energy $V(\mathbf{r})$, and marginal density $p_i(x_i) = \int p(\mathbf{x}) d\mathbf{x}_{j\neq i}$. Violations of Eq. (1) due to possible correlations are quantified by the well-known (Shannon) mutual information (MI),[22,15]

$$I[x_1, x_2, \ldots, x_{3N}] = \sum_{i=1}^{3N} H[x_i] - H[\mathbf{x}], \tag{2}$$

where $H[\mathbf{x}] = -\int p(\mathbf{x}) \log p(\mathbf{x}) d\mathbf{x}$ denotes the information-entropy. This nonlinear measure quantifies any correlation, that is, linear, nonlinear, and multicoordinate contributions. In the following we exploit this property by minimizing Eq. (2).

Full correlation analysis

We search an orthonormal coordinate transformation $\mathbf{R}$ of the Cartesian displacement vector $\mathbf{x}$ with $\mathbf{s}(t) = \mathbf{R}\mathbf{x}(t)$, such that $I[s_1, s_2, \ldots, s_{3N}]$ is minimal.

$\mathbf{R}$ is constructed iteratively by carrying out a sequence of rotations which respectively act on two coordinates x_i and x_j, respectively, that is, $\mathbf{R} = \prod_k \mathbf{R}_{i_k j_k}(\phi_k)$, where

$$\begin{aligned} &\mathbf{R}_{ij}(\phi) \cdot \left(x_1, \ldots, x_i, \ldots, x_j, \ldots, x_N\right)^{\mathrm{T}} \\ &\quad = \left(x_1, \ldots, \tilde{x}_i, \ldots, \tilde{x}_j, \ldots, x_N\right)^{\mathrm{T}}, \end{aligned} \tag{3}$$

with

$$\tilde{x}_i = x_i \cos\phi + x_j \sin\phi, \qquad \tilde{x}_j = -x_i \sin\phi + x_j \cos\phi.$$

A single plane rotation $\mathbf{R}_{ij}(\phi)$ leaves the two-dimensional information-entropy $H[x_i, x_j]$ invariant, such that the terms $H[x_i, x_j]$ and $H[\tilde{x}_i, \tilde{x}_j]$ cancel. Thus, MI changes by

$$\Delta_I(\phi) = I\left[\mathbf{R}_{ij}(\phi)\mathbf{x}\right] - I[\mathbf{x}] = H[\tilde{x}_i] + H\left[\tilde{x}_j\right] - H[x_i] - H\left[x_j\right]. \quad (4)$$

To find the global minimum of $\Delta_I(\phi)$ in a specific rotational plane, the angle ϕ was optimized in two steps. At first, the whole interval $\left[0, \frac{\pi}{2}\right]$ was sampled coarsely at 10 rotation angles $\{\phi_l\}_{l=1\ldots10}$. Subsequently, minimization was refined within the interval $[\phi_k - \Delta\phi, \phi_k + \Delta\phi]$, where ϕ_k is such that $\Delta_I(\phi_k)$ is minimal and $\Delta\phi$ denotes the step-size of the coarse sampling. The refinement used a combination of golden section search and parabolic interpolation, as implemented in the MATLAB™ fminbnd function.[24] The rotation planes for the minimization of MI were iterated with heuristically chosen coordinate pairs (i,j) until convergence. The algorithm is summarized by the following steps:

i. preprocessing: PCA to find an initial guess for **R**,
ii. heuristic choice of rotation plane (i,j),
iii. minimize $\Delta_I(\phi)$,
iv. repeat from (ii) until convergence.

Heuristic choice of rotation planes

For the high dimensional configuration space of proteins and other biological macromolecules the search space becomes too large, and randomly chosen rotation planes lead to slow convergence. We therefore developed heuristic pivot searches for the selection of planes for minimization of $\Delta_I(\phi)$ that had the largest potential to decrease MI. To this end, prior to every minimization step rotational planes (i,j) were ranked according to pairwise correlation $I_{ij} = I[\mathbf{x}_i, \mathbf{x}_j]$, since for large correlations a relatively high loss of mutual information upon minimization is expected. Furthermore, unnecessary reevaluations of already visited planes were avoided by using weights w_{ij} that were initialized with one, and set to zero after minimization in the (i,j)-plane. As rotation in the (i,j)-plane increased the likelihood that an already visited plane (i, k) or (j, k), $k \neq i,j$, allowed further optimization, all respective weights were increased by $|\phi|$, thus scheduling these planes for reevaluation.

Taken together, planes were evaluated in the order of decreasing $w_{i_1j_1}I_{i_1j_1} > w_{i_2j_2}I_{i_2j_2} > \ldots$, until four rotations with $|\phi| > 0.01$ were found. Then the pairwise correlations I_{ij} were recomputed and a new succession $w_{i_1j_1}I_{i_1j_1} > w_{i_2j_2}I_{i_2j_2} > \ldots$ was devised.

Note, that the estimation of pairwise correlations I_{ij} is computationally demanding. Because, I_{ij} was only used for the heuristics, small errors in I_{ij} were acceptable. Accordingly, only those pairwise correlations were updated, for which coordinates changed substantially. A second book-keeping matrix m_{ij} was used to track these changes. m_{ij} was set to zero after computation of I_{ij} and increased by $|\phi|$ if the rotation-angle corresponded to a (i,k) or (j,k) plane. For $m_{ij} > 0.3$ the respective correlation I_{ij} was re-evaluated. Convergence was assumed if all $w_{ij} \leq 0.01$.

Mutual information estimates

The FCA algorithm described earlier requires numerical estimates of entropies $H[x_i]$. Furthermore, the heuristic selection of rotational planes requires the explicit computation of pairwise mutual information $I[x_i, x_j]$, and hence estimates of $H[x_i, x_j]$. Accordingly, densities $p_i(x_i)$ and $p_{ij}(x_i, x_j)$ of one- or two-dimensional distributions, respectively, had to be estimated from the ensemble of coordinates x_i provided by the trajectory.

The choice of method was guided by conflicting criteria. On the one hand, the large number of evaluations during the iterative minimization requires a computationally efficient estimator. On the other hand, sufficiently high accuracy was required, because already small absolute errors in the information-entropy estimates $H[x_i]$ can cause large relative errors of their differences particularly if the difference is close to zero. Here we assumed that in typical applications of FCA the used MD ensembles of >10,000 structures were sufficiently large to render a fast kernel-smoothed histogram estimator preferable over more accurate but computationally much more demanding estimators such as spacing estimates,[25] k-nearest neighbor methods, or kernel density estimators.[26] To test this assumption, the accuracy of the fast kernel-smoothed histogram estimator was compared to an estimator based on a k-nearest neighbor approach (see Results Section).

The details of the required information-entropy estimation were as follows. The information-entropy $H[x_i]$ of a one-dimensional ensemble $\{x_i(t_k)\}_{k=1\ldots M}$ was estimated by counting occupations, n_b, of $b = 1 \cdots L_1$ bins, with $L_1 = 200$. The histogram was smoothed by convolution $p_b = \sum_{k=-m}^{m} n_{b+k}\, g_k/M$, with a discrete Gaussian function $g(l) = (2\pi\sigma^2)^{-1/2}\exp(-(l\Delta x^2/2\sigma^2))$ and the binning width $\sigma = \lambda_1\Delta_x$, evaluated at points $l = -m \cdots m$, with $m = 3$, and $\lambda_1 = 1$. From p_b the information-entropy of the ensemble $\{x_i(t_k)\}$ was computed according to $H[x_i] = -\Delta x\sum_{b=1}^{L} p_b \log p_b$.

Entropies of two-dimensional ensembles $H[x_i, x_j]$ were estimated by choosing $L_2 = 100$ bins for every dimension and widths $\sigma_i = \lambda_2\Delta x_i$ and $\sigma_j = \lambda_2\Delta x_j$, respectively, with $\lambda_2 = 1.8$. The reported parameter values L_1, L_2, λ_1, and λ_2 were empirically choosen to yield good estimates for Gaussian distributed data in the range of 10,000–100,000 structures.

For efficiency reasons we did not implement a sophisticated optimal bandwidth selection as, for example, in Ref. 27. A computationally less expensive bandwidth selection scheme[26] was tested, but led to unacceptable inaccuracies for distributions that deviated too much from a Gaussian. Instead, the bandwidth was chosen by

 DOI 10.1002/prot

adapting bin widths Δx_i and Δx_j such that a fixed number of bins (L_1 and L_2) covers the range between the extremes of the respective distribution. In this way, a satisfactory trade-off between efficiency, accuracy, and robustness, was achieved.

Preprocessing of FCA

Before minimization of MI commenced, PCA was applied to the C_α-atoms for the T4L example and to all non-hydrogen atoms of neurotensin, respectively. For efficiency reasons only rotations within the subspace of the first 100 eigenvectors were considered in both cases. The small amplitude PCA modes are already sufficiently uncoupled and, therefore, were not expected to change. This preprocessing step greatly enhances the efficiency, although it is not strictly required. In our experience a minimization starting from atomic coordinates yields similar FCA modes. Carrying out FCA on 100–200 degrees of freedom for ensembles of 10,000–30,000 protein structures, requires about 24–72 h on 6–10 of our dual Intel Xeon 3 GHz nodes of a Linux cluster.

Ranking of FCA modes

To select "essential" FCA modes, they need to be ranked. Rather than ranking by fluctuation amplitude $\langle x \rangle^2$ as for PCA,[9] we here ranked the FCA modes by anharmonicity. The anharmonicity of a collective mode was quantified by its negentropy,[18]

$$J[x_i] = \frac{1}{2}\left[1 + \log(2\pi) + \log\left(\langle x_i^2 \rangle\right)\right] - H[x_i], \qquad (5)$$

that is, the difference in the information-entropy of the observed density and that of a Gaussian function with the same variance. Note, that this definition, which we consider more appropriate for the present purpose, differs from the one given in Ref. 28. The mode with the highest anharmonicity was considered most essential, and denoted by the lowest index.

Selection of pairs of FCA modes

The subspace of relevant FCA or PCA modes is generally more than three-dimensional, and thus difficult to visualize. For exploratory data analysis and illustration, it is necessary to project the motion to pairs or triples of FCA modes, as it is customary for PCA modes.[9] However, the number of possible projections can grow quite large, and many projection pairs are redundant. The MI used for FCA offers the advantage to more systematically select pairs of modes that provide the best information, for example, a selection of modes with highest pairwise correlation. Accordingly, for the projections presented below, each of the first 10 modes was paired with that lower-indexed mode that showed the largest correlation to it.

A simple test-system

As a first test we constructed a set of 10 independent modes $s(t) = (s_1(t), s_2(t), \ldots, s_{10}(t))$. To mimic typical features of protein dynamics, five bimodal and five Gaussian distributions were generated with 30,000 points each. For the bimodal distributions, $s_i(t)$ $(i = 1 \cdots 5)$, five 300 ns trajectories were computed from a one-dimensional generalized Langevin model of the conformational motion of the peptide neurotensin, which was devised in Ref. 29. The five quasi-harmonic distributions, $s_i(t)$ $(i = 6 \cdots 10)$, were drawn randomly from Gaussian densities of differing widths $\langle x^2 \rangle^{1/2} = 1$, 0.8, 0.6, 0.4, and 0.2, respectively.

The 10 modes were mixed by applying a random orthonormal 10×10 matrix $\mathbf{A}$, with $\mathbf{x} = \mathbf{As}$. $\mathbf{A}$ was obtained from eigenvalue decomposition, that is, $\mathbf{TT}^{\mathrm{T}} = \mathbf{A}\Lambda\mathbf{A}^{\mathrm{T}}$, where $\mathbf{T}$ denotes a 10×20 matrix whose elements were drawn from normal distributed random numbers with unit variance, and Λ denotes a diagonal matrix.

From the mixed components $\mathbf{x} = \mathbf{As}$ recovery of the rotation matrix $\mathbf{R}$ was attempted by FCA and PCA, respectively. The accuracy of the obtained rotation matrices $\mathbf{R}_{\mathrm{FCA}}$ and $\mathbf{R}_{\mathrm{PCA}}$ was assessed by computing inner product matrices, $\mathbf{A}^{\mathrm{T}}\mathbf{R}_{\mathrm{FCA}}^{\mathrm{T}}$ and $\mathbf{A}^{\mathrm{T}}\mathbf{R}_{\mathrm{PCA}}^{\mathrm{T}}$, as well as recovered components, $\tilde{\mathbf{s}}_{\mathrm{FCA}} = \mathbf{R}_{\mathrm{FCA}}\mathbf{x}$ and $\tilde{\mathbf{s}}_{\mathrm{PCA}} = \mathbf{R}_{\mathrm{PCA}}\mathbf{x}$.

Collectivity of modes

We computed the collectivity Ω of a mode from its normalized direction vector in configurational space $\mathbf{d} = (d_1, d_2, \ldots, d_{3N})$, where N denotes the number of atoms, which is given as a column of the product matrix $\mathbf{R}_{\mathrm{FCA}}\mathbf{R}_{\mathrm{PRE}}$, where $\mathbf{R}_{\mathrm{PRE}}$ denotes the $100 \times 3N$ matrix gained with PCA in the preprocessing step (see earlier). To this end, the squared contribution a_i^2 of the fluctuation of atom i to mode s was computed as the sum of the squared components that belong to atom i, i.e., $a_i^2 = \sum_{j=1}^{3} d_{3(i-1)+j}^2$. The collectivity was given by the information-entropy of the distribution of motional contributions,

$$\Omega(\mathbf{d}) = -\frac{1}{\log N} \sum_{i=1}^{N} a_i^2 \log a_i^2.$$

The normalization constant $\log N$ was chosen such that for a mode to which all atoms contributed equally a collectivity of one is regained.

The generalized correlation coefficient

To characterize the remaining coupling between FCA modes, we used the pairwise mutual information $I[\mathbf{x}_i, \mathbf{x}_j]$,

which can be computed explicitly, in contrast to the full MI of all $3N$ degrees of freedom, Eq. (2).

However, I yields values in the range $[0\cdots\infty)$, which is unfamiliar and has no obvious interpretation. On the contrary, the magnitude of the well-known linear Pearson correlation coefficient $r = |\langle x_i x_j\rangle/(\langle x_i^2\rangle\langle x_j^2\rangle)^{1/2}|$, has the familiar interpretation with $r = 1$ fully correlated and $r = 0$ uncorrelated, respectively. To allow a similarly intuitive interpretation also for I, we used the previously suggested generalized correlation coefficient, r_{MI},

$$r_{\mathrm{MI}}\left[\mathbf{x}_i, \mathbf{x}_j\right] = \left(1 - e^{-2I[\mathbf{x}_i,\mathbf{x}_j]/d}\right)^{\frac{1}{2}}, \qquad (6)$$

which yields—in case of purely linear correlations—the same result as the Pearson coefficient, but also captures nonlinear correlations,[22] because it is derived from I.

Molecular dynamics simulations

A 117 ns molecular dynamics (MD) simulation, T4L, was started from the crystal structure of coliphage T4 lysozyme M6I (PDB entry 150L chain D). The protein was solvated in 8898 TIP4P water molecules and 8 Cl^- counter ions using a rectangular box.

For a second simulation neurotensin (a peptide with the sequence Ac-RRPYIL[30]), was solvated with 2246 TIP4P water molecules and 2 Cl^- counter ions in a cubic box. A 100 ns simulation was started from an extended configuration of the peptide.

All MD simulations were carried out using the Gromacs simulation suite[31] together with the OPLS all atom force field.[32] Lincs and Settle[33,34] were applied to constrain covalent bond lengths, allowing an integration time-step of 2 fs. Electrostatic interactions were calculated using the Particle-Mesh-Ewald method.[35,36] The temperature was kept constant by separately coupling the peptide and solvent to an external temperature bath (τ = 0.1 ps).[37] The pressure was kept constant by weak isotropic coupling to a pressure bath (τ = 0.1 ps).[37] Prior to analysis, all recorded structures were superimposed to the crystal structure as a reference.

Conformational transition of neurotensin

Calculation of free energy surfaces $G(s_1, s_2) = \beta^{-1}$ log $\rho(s_1, s_2)$ on a subspace spanned by pairs of modes (s_1, s_2) required determination of the density $\rho(s_1, s_2)$ of the projected MD ensemble. This density was estimated by smoothing a two-dimensional histogram (150 × 150 bins) with a Gaussian function of widths $\sigma_1 = 3\Delta s_1$ and $\sigma_2 = 3\Delta s_2$, respectively, where Δs_1 and Δs_2 denote the bin widths. The superposed trajectories shown in Figure 8 were obtained by projecting the MD trajectory onto the respective modes and subsequent smoothing with a Gaussian function of width σ = 20 ps.

RESULTS AND DISCUSSION

Accuracy of entropy estimates

For efficiency reasons the MI was calculated from a relatively crude but fast estimator based on histograms. To evaluate the accuracy of this approach estimates were compared with those obtained from the recently devised k-nearest neighbor approach of Kraskov et al. as a reference. This estimator is unbiased and was found to be more accurate than a number of other methods.[38] To this end, MI estimated from Gaussian distributions with random widths were compared to MI calculated analytically.

Estimates of entropies determined via a histogram generally depend on the chosen bandwidth, which is given here by the size of the bins. The optimal bandwidth depends on the statistics of the data, and usually this optimum even shifts in dependence on the number of sample points. In our approach the bandwidth was determined implicitly by using a fixed relation of bandwidth to bin-size (λ_1,λ_2) and by using a given number of equidistant bins for the range spanned by the data points (see Methods). Being aware of a possible shift in optimal bandwidth, we checked whether the implicitly chosen bandwidth yields sufficient accuracy for both boundaries of the envisaged range of $M = 10^4\cdots10^5$ of sample points.

In Figure 1 MI estimated with both, the histogram method and Kraskov's method, are plotted against the analytically obtained MI for $M = 2 \times 10^4$ and $M = 10^5$ (cf. inset) sample points. As can be seen, estimates from the histogram method were not less accurate than those

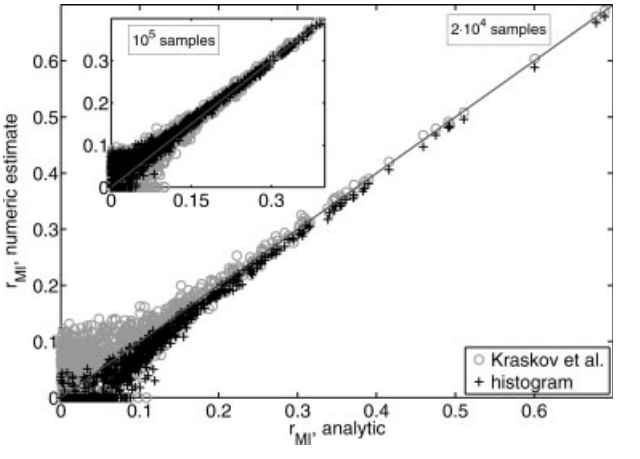

Figure 1

Estimation of correlation for Gaussian distributed random data sets. r_{MI} values estimated with the histogram method (crosses) and with the method of Kraskov et al.[38] (circles) is plotted against analytically computed r_{MI}. The inset shows the same comparison, but with 10^5 sample points used.

DOI 10.1002/prot

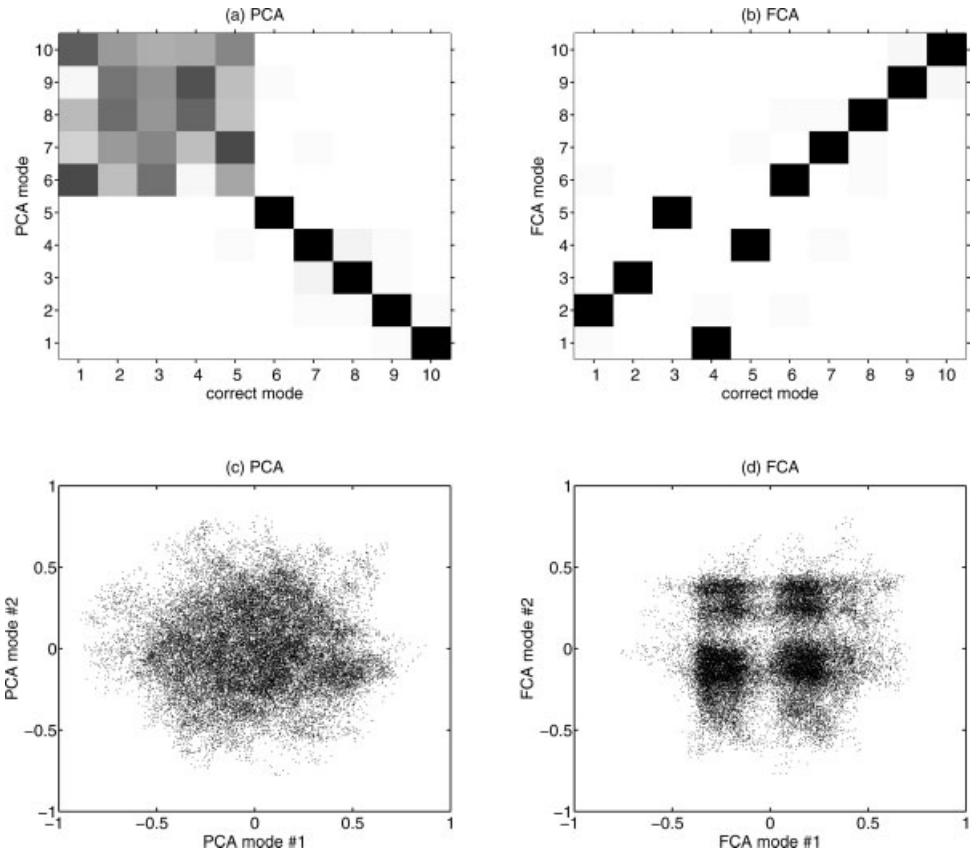

Figure 2

*(**a**,**b**) Inner products of PCA (a) and FCA (b) modes with the directions of the independent components s_i. The black squares denote inner products of near unity. (**c**,**d**) Projections of the test-ensemble onto the first two modes calculated from PCA (c) and FCA (d), respectively.*

of the reference method. In particular, in the low correlation regime the histogram estimates were even more accurate than Kraskov's estimates.

Above we have shown that the histogram based approach obtains accurate estimates of MI for Gaussian distributions. To rule out a significantly lower accuracy for non-Gaussian distributions, we checked also MIs of distributions obtained from molecular dynamics data of T4 lysozyme using Kraskov's method as a reference (results not shown). The achieved correlation with the reference (correlation coefficient $r = 0.98$) shows that for MD ensembles the histogram method reaches nearly the accuracy of the computationally much more expensive method. Therefore, we chose the faster histogram method for FCA.

We note that a very recently developed method[39] offers high accuracy at computational costs comparable to the histogram method, but had not yet been available at the time when the presented work was performed. Meanwhile, we have, however, implemented the newer approach into our FCA software.[40]

Application of FCA to a mock-protein ensemble with known result

As a first test, we applied FCA and PCA to a synthetic example where we already knew the coordinate transformation $\mathbf{A}^T$ that transforms the atomic coordinates $\mathbf{x}$ of a mock-protein ensemble into uncoupled modes $s_1(t)$, $s_2(t), \ldots, s_{10}(t)$.

To quantify the results, note that the directions of the original components s_i in the atomic coordinate system $\mathbf{x}$ are given by the columns of $\mathbf{A}$. Hence, an accurately identified mode would yield an inner product near unity with exactly one of the columns of $\mathbf{A}$. Figure 2(**a**) shows the respective inner products for PCA modes. Apparently, the field of gray boxes in the upper left shows that PCA was not able to recover the anharmonic modes $s_1, s_2, \ldots, s_5$, whereas the black boxes in the lower right demonstrates that the quasi-harmonic modes $s_6, \ldots, s_{10}$ were retrieved successfully. For FCA, encouragingly, all 10 independent components were accurately recovered, as shown by 10 inner products near unity [black boxes, Fig. 2(b)].

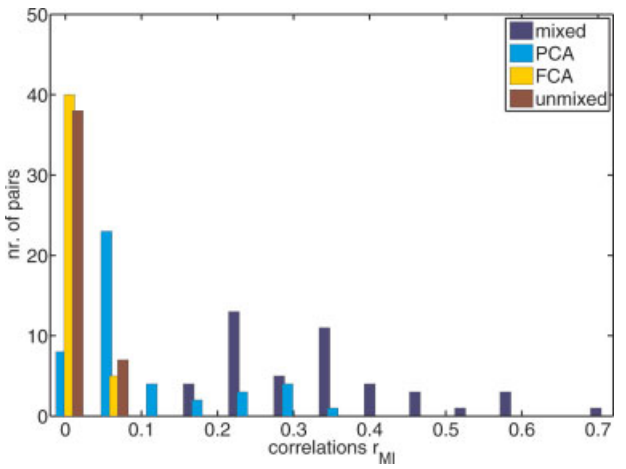

Figure 3

Histogram of generalized correlation coefficients between pairs of coordinates. The four histograms count correlations between the s_i (unmixed), the x_i (mixed), and between PCA and FCA modes, respectively.

As a consequence, the projection to the first two FCA modes shown in Figure 2(d) revealed correctly the peaked structure of the conformational density, whereas this structure is completely obscured in the projection to the first two PCA modes [Fig. 2(c)], as it is the case for any other combination of PCA modes (not shown).

FCA aims at finding collective coordinates that are the least correlated. To assess all pairwise correlations, Figure 3 displays histograms of all off-diagonal matrix elements of the generalized correlation matrix $r_{MI}[c_i, c_j]$, with c_i denoting either the pseudo-atomic coordinates x_i, the PCA or FCA modes, or the original input coordinates s_i, respectively. As can be seen the pairwise correlations were significantly reduced for pairs of FCA modes compared to PCA modes. Small correlations below 0.1 between FCA modes remain, which are due to the finite number (30,000) of sample points and statistical inaccuracies in their estimation (cf. Fig. 1). Accordingly, these remaining correlations also appear between pairs s_i,s_j of the input modes, which are uncorrelated by construction. As expected, PCA achieved only partial reduction of the correlations.

In this test example the algorithm clearly reached the overall minimum of mutual information. However, it should be stressed that there is no guarantee that the global minimum is always found. Mutual information depends nonlinearly on the chosen directions of the FCA-modes, such that it is likely that also local minima are visited by the FCA algorithm. Choosing the global minimum for every single rotation plane allows the algorithm to leave most local minima, although it may get stuck in a local minimum if concerted rotations in multiple planes are required to reach lower values of mutual information. We should point out, however, that the (known) global minimum for the presented test-case (cf. Fig. 2) was always found independently of various starting conditions. Moreover, our experience showed that the minimization always proceeded sufficiently far to yield collective coordinates of similar beneficial characteristics. Nevertheless, further work should address this issue.

Whereas this first test case was quite illuminating, it lacked the likely property of protein dynamics that there are no fully uncoupled modes. Therefore, we tested whether FCA is able to reverse the mixing also in cases where the known solution contains coupled modes. Indeed, FCA also solved such test examples that were constructed to contain pairs and triples of coupled coordinates (results not shown).

This further test directly bears on the ability of FCA to handle nonlinear motions. For PCA separation of nonlinear motion cannot be achieved unless nonlinear coordinates are considered, which, however, involves considerable technical and conceptual challenges.[41–43] Also FCA considers linear (orthonormal) transformations of the atomic Cartesian coordinates. Thus, uncorrelated curvilinear motions will not be represented by single FCA modes. However, in contrast to PCA, FCA does separate also curvilinear motions into blocks of modes which have high intra- but no(low) inter-block correlation. Consider, for example, two uncorrelated circular motions. FCA will describe these using four FCA modes. Each circular motion represented by a block of two highly correlated FCA modes (i.e., the sin and cos component of its phase), whereas there is no correlation between modes from different blocks. As reported earlier, we successfully tested the algorithm on mock-protein ensembles that contained pairs or triples of coupled coordinates, which resulted in correlated blocks of FCA modes. Moreover, this block-structure due to nonlinear motion is observable in one of the real-world examples presented later [cf. Fig. 15(b), e.g., modes 1/2, 26/27, 28/29]. Nevertheless, it might be worth-while to integrate the recent nonlinear coordinate approaches into the framework of FCA, although their benefit is here likely to be less pronounced than for PCA.

Before proceeding to the application of FCA to real proteins, we briefly discuss the relation of the FCA algorithm to algorithms used for the related Independent Component Analysis (ICA) known from signal processing (cf. Introduction).

The aim of ICA is to recover the underlying independent sources from a recorded multichannel signal of their observed mixture. Within the context of molecular simulations, the Cartesian coordinates represent observation channels, and the collective motions are the putatively independent signals supposed to be recovered. ICA algorithms usually simplify the search problem by applying the so-called pre-whitening, that is, a scaling which imposes unity on all eigenvalues of the covariance

 DOI 10.1002/prot

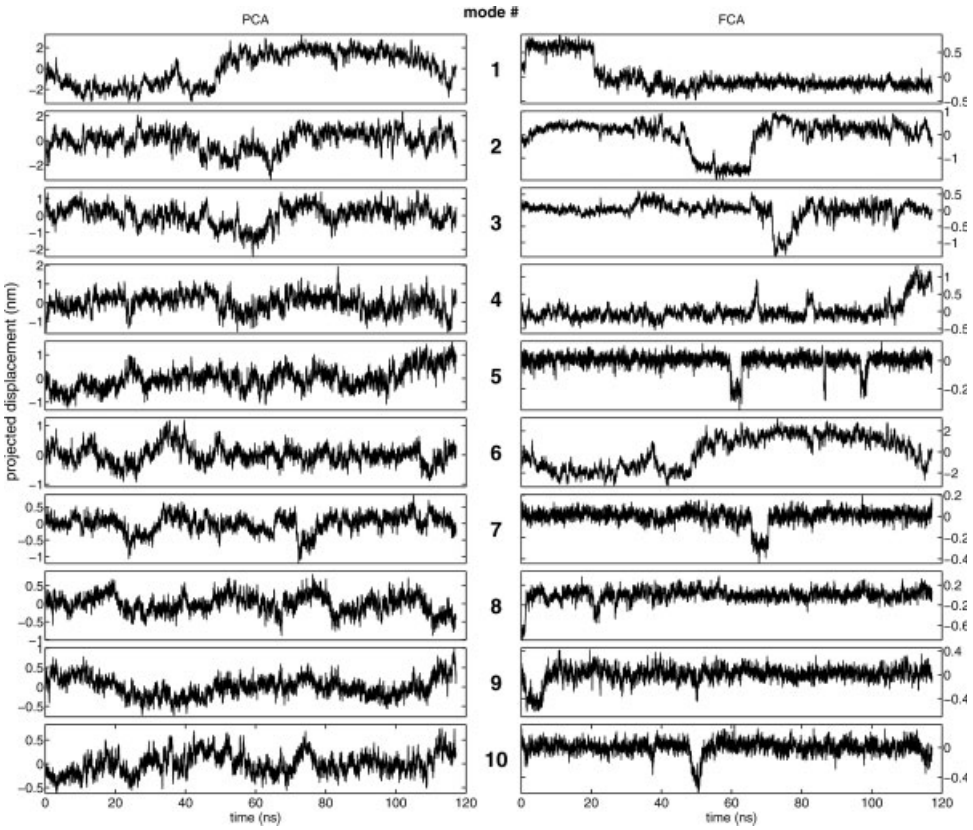

Figure 4

Projections of T4L MD simulation trajectories onto the first 10 PCA (left) and FCA modes (right), respectively.

matrix[19,44,23] and thereby remove possible scaling bias. On the contrary, for protein dynamics the relative amplitudes of all coordinates are important. Therefore, the simplification offered by pre-whitening was not applied here. Moreover, in analogy to PCA, FCA was set-up to be restricted to rotations in configurational space, thereby conserving the geometry of the conformational ensemble. One particular purpose of this restriction was to keep phase space volumes unchanged, thus enabling a straightforward computation of thermodynamics quantities such as free energies or entropies.

The algorithm devised for FCA is an adaption of MILCA, which outperforms many other ICA algorithms.[20] The main changes upon MILCA lie in the treatment of MI. For FCA, the sum of single dimensional entropies, Eq. (4), was minimized directly, whereas MILCA minimizes pairwise MI. At first glance, this is equivalent, that is, in analogy to Eq. (4) MILCA uses

$$\Delta_I(\phi) = I\left[\mathbf{R}_{ij}(\phi)\mathbf{x}\right] - I[\mathbf{x}] = I\left[\tilde{x}_i, \tilde{x}_j\right] - I\left[x_i, x_j\right]. \quad (7)$$

However, this implicitly [cf. Eq. (2)] involves estimation of two-dimensional entropies $H[x_i,x_j]$ and $H[\tilde{x}_i,\tilde{x}_j]$, which renders $\Delta_I(\phi)$ prone to statistical errors because of the near cancellation of large information-entropy values (as discussed earlier). Because of these inaccuracies, the right hand side of Eq. (7) is a highly rugged function, such that identification of the global minimum proves difficult. Applying here Eq. (2) instead of Eq. (7) renders $\Delta_I(\phi)$ much smoother such that we require about a tenth of the evaluations of $\Delta_I(\phi)$ in a single rotation plane.

As a further difference to MILCA rotational planes were here chosen systematically, which increased convergence speed.

Conformational motion of lysozyme analyzed with FCA

Having provided evidence that the FCA algorithm works as intended we applied FCA to a real protein system, T4 lysozyme. We have chosen this protein because it exhibits pronounced domain motion, which is essential for function of T4L allowing the substrate to enter and the products to leave the active site.[45–48] Accordingly, the ensemble of T4L structures gained from a 117 ns MD

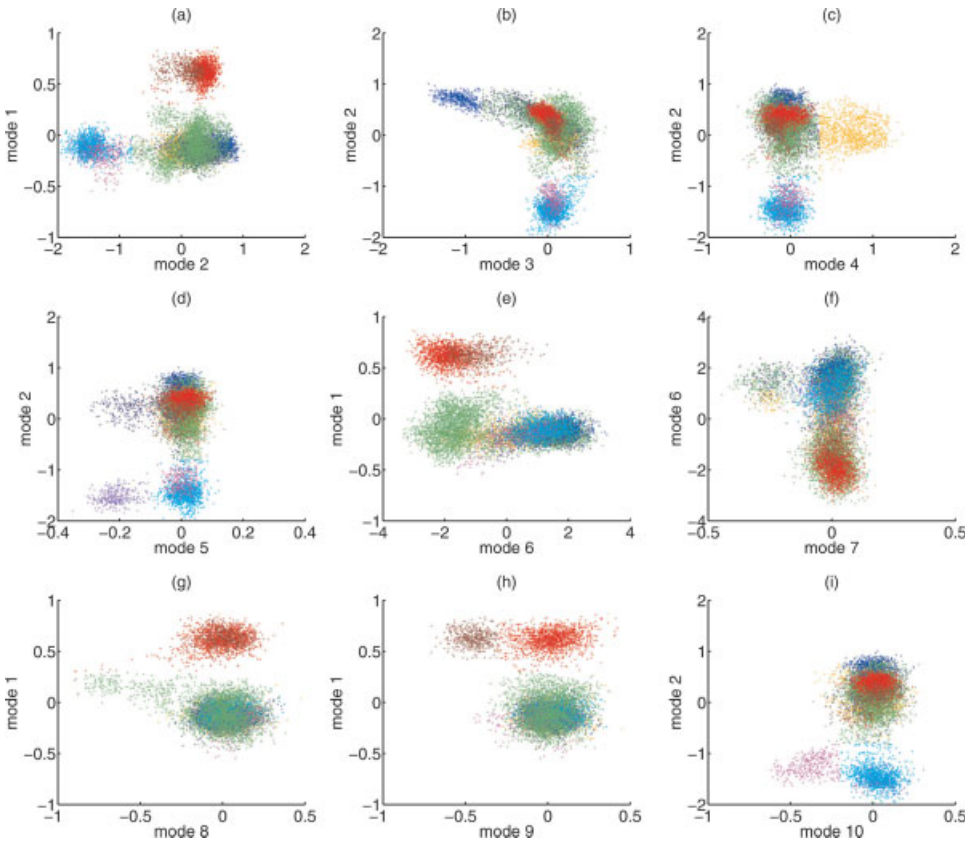

Figure 5

Projections of a 117 ns T4L MD simulation trajectory of T4L onto pairs of FCA modes. The presented pairs of FCA modes were selected based on pair correlations and anharmonicity, as described in Methods. The temporal sequence of frames is color coded (see-text).

simulation was analyzed with FCA and, for comparison, also with PCA.

Figure 4 shows 10 projections onto PCA (left) and FCA (right) modes as a function of time. As can be seen virtually all projections to FCA modes show pronounced differences from those to PCA modes, with the single exception that PCA mode 1 is nearly identical to FCA mode 6. Overall, and in contrast to PCA modes, the fluctuations in the FCA modes are relatively small, only interrupted by larger transitions.

We note that in Figure 4 PCA modes were sorted by fluctuation amplitude, whereas FCA modes were sorted by anharmonicity (cf. Methods). We consider such direct comparison of the differently ranked modes justified, since the ranking scheme is an essential part of the respective methods. Nevertheless, neither in the highly anharmonic nor in the large amplitude PCA modes, transitions were as clearly distinguishable from the background fluctuations as in PCA mode 1 or FCA modes 1–10.

This crucial feature becomes even more apparent when turning to projections of the MD ensemble of T4L onto pairs of PCA and FCA modes. This type of projections is often used for analysis of conformational states and transitions between them, because it reveals conformational states as clusters of points. Figure 5 shows projections of the MD ensemble of T4L onto those pairs of FCA modes that were selected based on correlation and anharmonicity using the protocol described in Methods. As can be seen in the presented projections, the colored points, which each represents a particular structure of T4L, cluster into many different conformational substates. In Panel (a), for instance, three clearly separated clusters are visible. Note that the color chosen for the different frames is the same in all panels to allow re-identification of configurations in the different projections. For example, in Panel (b) the blue points with very low values of FCA mode 3 constitute a different conformational substate than the red points separated off in Panel (a) by mode 1. Similarly, by using all presented projections, we assigned a different color to all identified conformational substates.

Strikingly, the projections to pairs of FCA-modes often adopted an *L*-shape. Thus, FCA tended to describe tran-

DOI 10.1002/prot

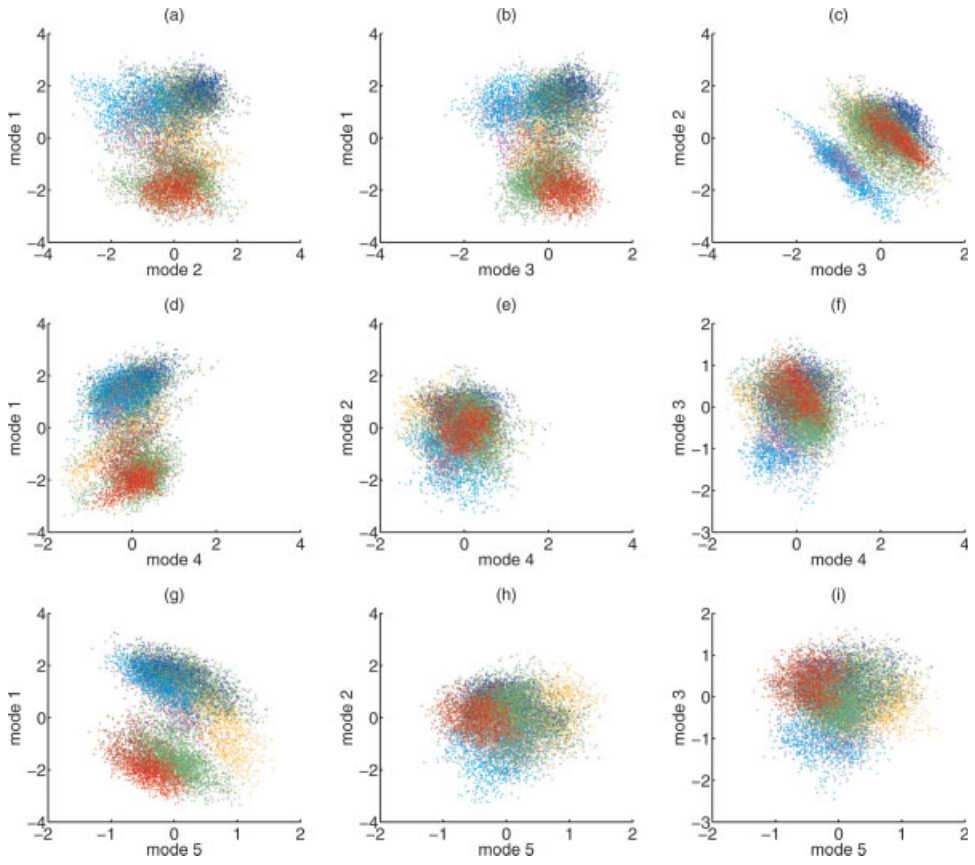

Figure 6

Projections of a 117 ns T4L MD simulation trajectory onto pairs of PCA modes. The coloring of the points corresponds to that in Figure 5.

sitions between two conformational substates with a single FCA mode. For instance, mode 2 described a transition in Panel (a) from the center cluster (blue, green, and yellow points) towards the cluster on the left (cyan and magenta points). Following this mode through all panels shows that mainly two conformations, that is, the cyan cluster and the green cluster, were involved in this transition, since all other conformations (blue, yellow, and red) had interconverted with the green conformations via modes 3 (Panel b), 4 (Panel c), and 1 (Panel a), respectively, before they underwent transitions to the cyan conformations via mode 2. Possibly also the magenta conformations (Panel i) were involved in the cyan–green interconversion. Taken by itself, however, the projection shown in Panel (i) of Figure 5 did not allow to decide whether the magenta conformations were reached from the cyan conformations after the transition along mode 2 had taken place or as an intermediate substate during the transition. As revealed by the time information of the FCA modes shown in Figure 4, the latter is the case. The magenta conformations were visited as intermediate substate (48–52 ns, Fig. 4) during the course of the main transition along mode 2 (48–56 ns, cf. Fig. 4). Accordingly, the simultaneous motion along the two modes shows up in Figure 5(i) as a diagonal connection between the green and magenta clusters. The intermediate substate was not visited during the back-transition to the green conformations (65–65.5 ns, cf. Fig. 4). Hence, FCA tended indeed to describe transitions between major conformational states by single modes, whereas minor intermediate conformational substates, that is, a more detailed picture of the transition pathway, became only resolved with additional FCA modes.

For comparison, Figure 6 shows projections of the ensemble to pairs of the first five PCA-modes. From projections (a, b, c, d, and g), also a clustering into two or three conformational substates may be inferred, albeit much less resolved, as observed previously.[49] To reidentify the conformational substates revealed previously by FCA in the projections to the PCA modes, the same color-code as in Figure 5 was used in this plot. Apparently, most conformational substates revealed by FCA overlapped strongly in the projections to PCA modes, such that only in the projection to mode-pairs 3:2 and

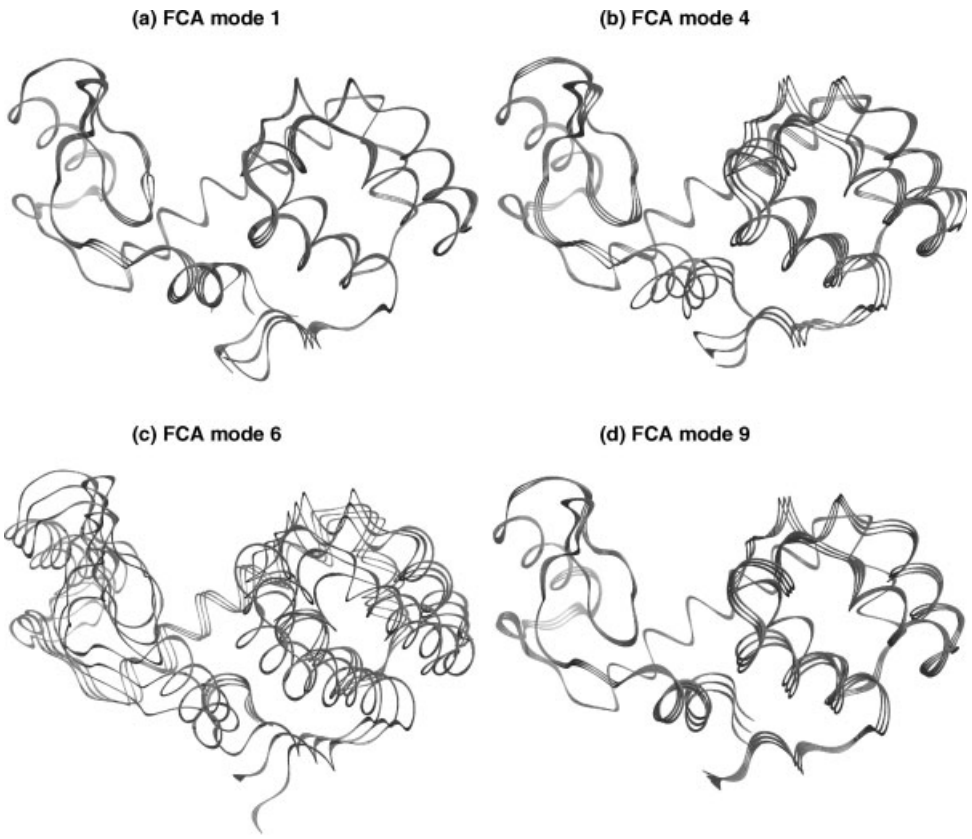

Figure 7
Superposition of three T4 lysozyme configurations obtained by projecting its C_α motion of T4 lysozyme onto the respective FCA mode.

5:1 [Fig. 6(c,g), respectively] an assignment of points to their respective clusters would have been unambiguous. Nevertheless, as seen from the colors, also in these projections several different conformational substates would be assigned to the same super-state, with insufficient resolution to reveal the finer substructure.

Moreover, Figure 6 indicates that the tendency of FCA to align its modes with actual conformational transitions—that is, to uncouple these transitions—is not shared by PCA. For example, the transition along FCA mode 2 was described by PCA modes 2 and 3, and to a lesser extent also by PCA mode 4. Consequently, motions which do not contribute to the transition were also mapped onto these PCA modes, thereby, causing their large fluctuations during the whole simulation length (cf. Fig. 4).

To visualize which motions are actually described by the obtained FCA modes, Figure 7 shows superpositions of three structures obtained by projecting the C_α motion of T4 lysozyme onto four selected FCA modes. FCA mode 1 [cf. Fig. 7(a)] corresponds to a local swiveling motion of the 3 N-terminal residues, whereas FCA modes 2 and 3 describe a similar motion of the C-terminus (not shown). FCA mode 4 and 9 [Fig. 7(b,d)], describe collective motions involving the whole C-terminal domain and helix 1. FCA mode 6 [Fig. 7(c)]—as well as the identical PCA mode 1—show a highly collective motion of the whole protein.

FCA modes 6 and 21 (not shown) describe the previously identified closure and twist motion of the two domains relative to each other.[49] FCA modes 4 and 9, reveal more intricate details of the dynamics. Along FCA mode 4 a bundle of the four parallel helices in domain 2 (H5, H7, H8, and H10) rotates inwards pushing the functionally important H1 such that its axis tilts outwards. FCA mode 9 reveals an opening of domain 2 by moving H9 outwards that is correlated to a shift of the inner plane constituted by helices H1, H5, H6, and H7 against the outer plane of H8—H10.

The presented projections of the T4L ensemble to FCA modes showed a substantially improved resolution of conformational substates as compared to PCA modes. Furthermore, transitions between substates are described by single FCA modes, suggesting that FCA is particularly

 DOI 10.1002/prot

suitable to yield optimally uncoupled conformational coordinates (reaction coordinates), which is an indispensable prerequisite for dimension reduced approaches. The following subsection further explores the suitability of FCA for dimension reduced descriptions.

Conformational transitions of neurotensin with FCA

For a dimension reduced description of conformational dynamics, free energy surfaces of low dimensional subspaces spanned by (collective) degrees of freedom need to be computed. In most cases chemical and/or physical intuition is used for selection of collective coordinates (also named order parameters), such as center of mass distances, approaching angles, radius of gyration, and so forth. However, the particular choice of coordinates critically determines the quality of the dimension reduced description, and unsuitably chosen coordinates can lead to wrong reaction pathways, barrier heights, and transition rates. The main cause of such artifacts are conformational substates that are well separated in full configurational space, but projected such that they overlap in the subspace, and, therefore, their separating free energy barrier seems too low. Therefore, any improvement in the ability to separate substates also implies improved description of the dynamics within the subspace.

As discussed previously, both, PCA and FCA, enable a systematic selection of suitable collective degrees of freedom for such a reduced description. As shown in Figures 5 and 6 above, FCA modes yielded an improved resolution of the substates, and less modes were needed to describe a conformational transition. In the following comparison, we focus on the quality of free energy surfaces spanned by selected pairs of modes. As an illustration, we will compare the pathways of transitions actually observed by MD with pathways obtained from the topography of the free energy surfaces.

Figure 8(a) shows free energy surfaces of two PCA modes derived from a 100 ns MD simulation of neurotensin. Two major conformational states, denoted *A* and *B*, were identified as two shallow basins of the free energy surface (blue). The two minima are connected by a channel of relatively low free energy, implying a putative transition state at $(s_1, s_2) \approx (-0.5, -1)$ (cross), which is ≈ 1 $k_B T$ lower in energy than the remaining transition region (white bar). One would, therefore, expect to find most transitions between the two conformational states to proceed through this channel. However, quite the contrary was observed: As can be seen from the smoothed trajectory (black), all successful transitions occurred via the region between $-0.5 < s_2 < 1.5$ (white bar), where the energy is about 1 $k_B T$ higher than the suggested transition channel. In fact, only two (unsuccessful) crossing attempts (arrows) explored the putative lowest free energy path, with subsequent immediate return to state B. This peculiar behavior was explained by including more PCA modes into the analysis. Rather than reaching state *A*, the system remained in a protrusion of conformational state *B*, whose projection just happened to overlap with the projection of conformational state *A* (data shown in Ref. 29). It is this overlap that caused by

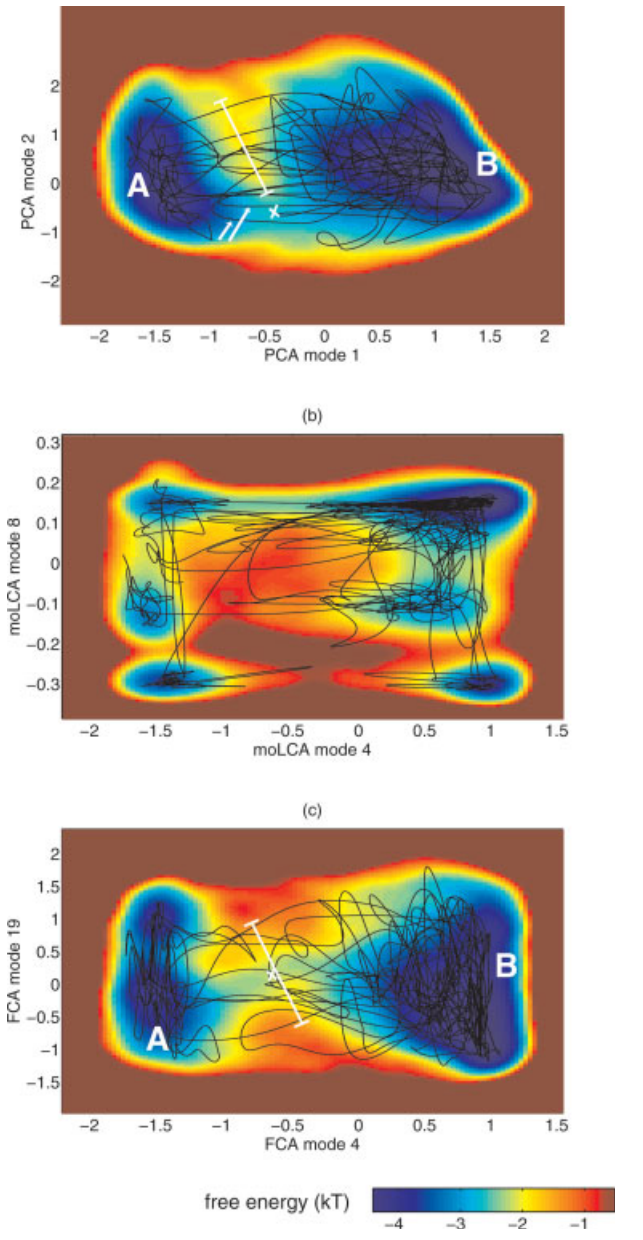

Figure 8

*Free energy surfaces derived from neurotensin MD trajectories: the smoothed projected trajectories are plotted (black) on top of the free energy surface (colors) in projections to (**a**) two PCA and (**b,c**) two FCA modes, respectively. Smoothing with a Gaussian kernel function (filter width 10–20 ps) suppresses intra substate fluctuations and reveals the transitions more clearly. The two arrows in (a) denote two unsuccessful transition attempts, which due to a projection artifact create the wrong impression of a transition from state B to state A, see text.*

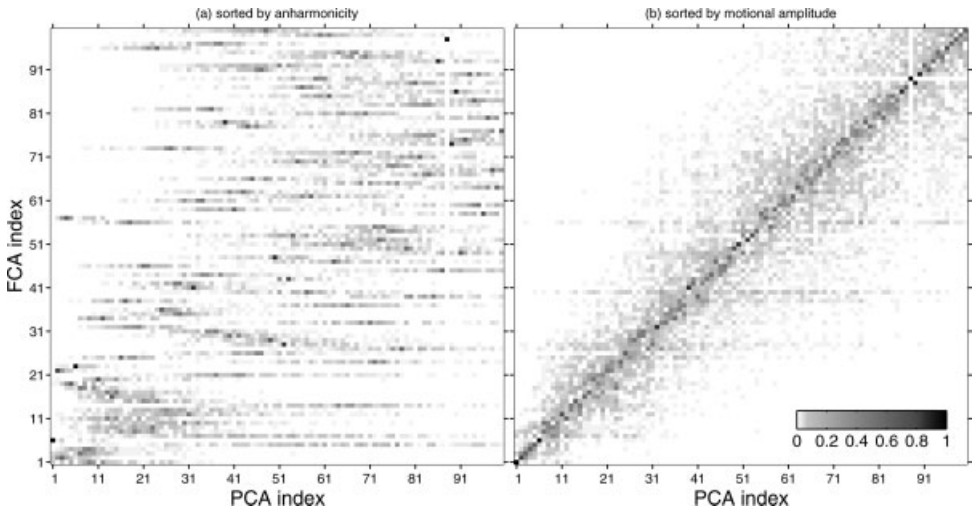

Figure 9

*Inner product matrices between FCA and PCA modes of T4 lysozyme, sorted by (**a**) degree of anharmonicity and (**b**) fluctuation amplitude.*

inappropriate projection the misleading low free energy channel.

Can FCA provide an improved energy landscape? Figure 8(b) shows pair number four (modes 4 and 8) from the first nine automatically selected pairs (cf. Methods) that showed the most pronounced clustering. In this FCA projection each of the conformational states *A* and *B* separates into three subclusters, and the transition pathways agree well to the low free energy valleys. But even if those FCA modes were selected that are most similar to PCA modes 1 and 2 [Fig. 8(c)] an improved free energy landscape is obtained. As in Figure 8(a), two conformational states are resolved, but here the channel of lowest free energy does agree with the observed transition pathways. Moreover, FCA mode 19 revealed a substructure of conformational state A, which was not resolved by the PCA modes.

The presented results suggest that FCA modes cause less projection artifacts than PCA modes. In this sense, the collective coordinates extracted by FCA render protein conformational dynamics better accessible to dimension reduced dynamics.

Comparative analysis of PCA and FCA modes

The previous sections have shown distinctly different characteristics between projections of protein dynamics onto FCA modes and onto PCA modes. To pin-point the origin of these changes, we will subsequently characterize the differences between FCA and PCA modes.

First, the directions of FCA and PCA modes were compared. To this end, their mutual colinearity was quantified by inner products depicted in Figure 9 for T4 lysozyme and in Figure 10 for neurotensin. On the left hand side of both figures, the FCA modes are sorted by their degree of anharmonicity as defined and used in the previous sections; on the right hand side they were sorted by fluctuation amplitude. The figures show that ordering of FCA modes by anharmonicity prevents a direct comparison with PCA modes. Therefore, and to avoid any confusions, we will sort in the following both, FCA and PCA modes, by fluctuation amplitude.

For T4 lysozyme, Figure 9(b) shows that almost all FCA modes differ from PCA modes (the maximum inner product with a PCA mode is below 0.9). In particular, from the low indexed FCA modes only the directions of mode 1 (previously mode 6) and mode 6 (previously 23) are colinear to PCA modes. Nonetheless, it is evident that specific FCA modes are generally contained in a low-dimensional subspace spanned by PCA modes of similar amplitude. For instance, many PCA modes below 30 contribute to FCA mode 7, and FCA mode 50 is a combination of PCA modes between 30 and 80. Thus, the PCA and FCA subspaces of large amplitude modes of T4L overlap to a large extent, although the directions of their respective basis vectors differ. Note that this finding justifies the restriction of the FCA minimization on a sufficiently large subspace spanned by PCA modes instead of a direct minimization of all atomic coordinates (see Methods).

For neurotensin (NT), FCA modes were generally less colinear with PCA modes than for T4L [cf. Fig. 10(b)].

DOI 10.1002/prot

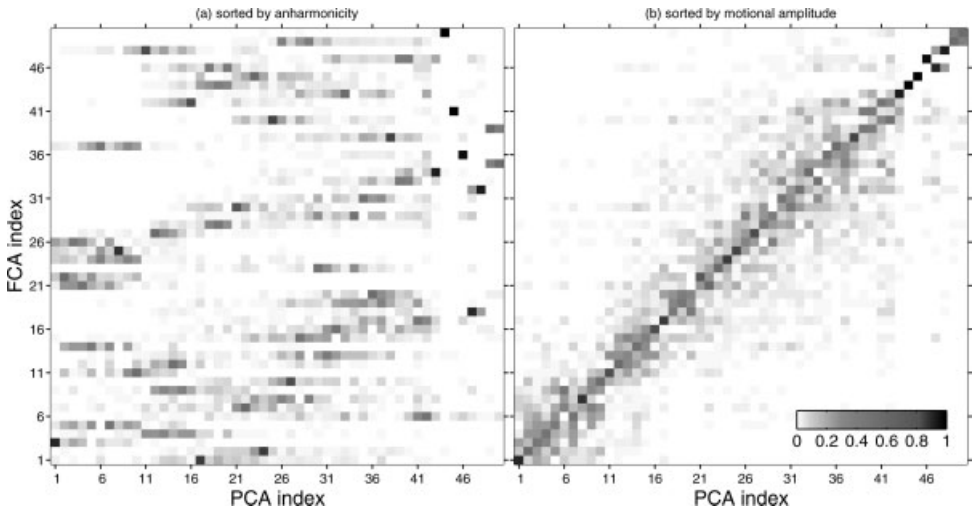

Figure 10

Inner products between FCA and PCA modes of neurotensin. In the left plot the FCA modes are sorted by anharmonicity and in the right by fluctuation amplitude.

However, only the first 10 PCA modes contributed significantly to the first 10 FCA modes, that is, the large amplitude subspaces overlapped, as observed already for T4L.

An important and often exploited property of principal components of protein ensembles is the fast decrease of their fluctuation amplitude. Figure 11 shows that for both test systems the fluctuation amplitude of FCA modes does not differ significantly from that of the PCA modes, although FCA optimizes mutual information instead of the fluctuation amplitude. Therefore, the often very useful property of PCA that the first few modes describe a major part of the total atomic displacement of the protein ensemble,[9] is shared by FCA.

Aiming at functionally relevant motions one is generally not interested in extracting modes that describe very local motions, for example, displacement of single C_α-atoms or the flip of single side chain dihedrals. Low-indexed PCA modes are known to be typically highly collective, since they maximize the fluctuation amplitude, which is generally the larger the more atoms are involved. FCA, on the contrary, has a less direct link to collectivity. To address this aspect, Figure 12(a) compares the collectivity of FCA and PCA modes. As can be seen three FCA modes of T4L exhibit indeed a relatively little collectivity. These modes describe the swiveling motion of either 3 C-terminal or 3 N-terminal residues [cf. Fig. 7(a)]. The two PCA modes with lowest collectivity similarly described such swiveling motions of the terminal residues, but were less focused at it, such that their collectivity was slightly higher than that of their FCA counterparts. All other FCA modes had a collectivity similar as PCA modes. For NT, the collectivities of FCA and PCA were also very similar [cf. Fig. 12(b)]. Unexpectedly, here the most localized modes were obtained by PCA.

So far we have shown that, amplitudes and collectivities of FCA and PCA modes did not differ very much. In harsh contrast, Figure 13 reveals large differences for the anharmonicities, Eq. (5). The shown scatter plot of anharmonicity and collectivity reveals that for both test systems, T4L and neurotensin, only FCA combined high collectivity and high anharmonicity. In particular, for

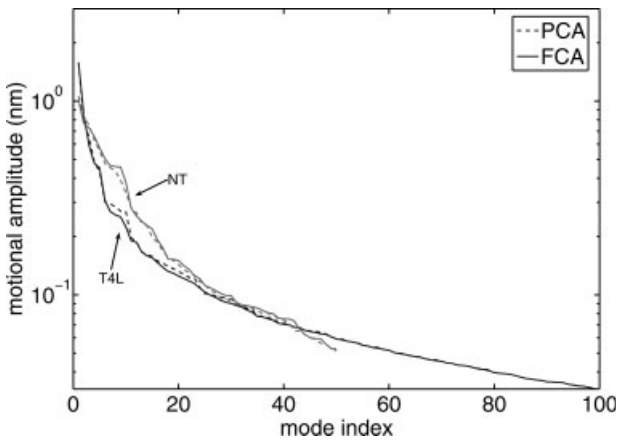

Figure 11

Fluctuation amplitude of PCA and FCA modes of T4 lysozyme (T4L) and neurotensin (NT). FCA modes are sorted by fluctuation amplitude.

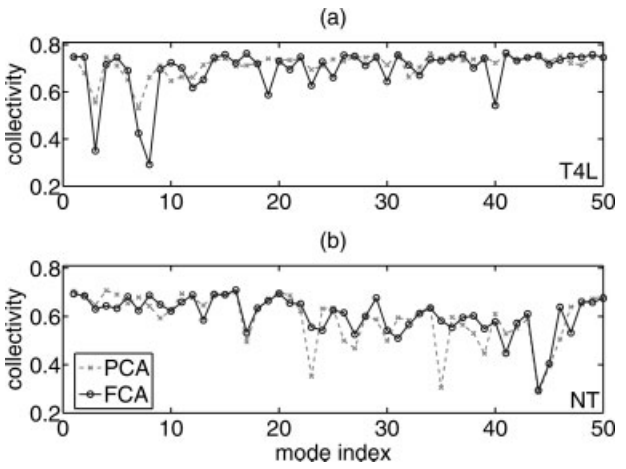

Figure 12

*Collectivity of the motion described by PCA and FCA modes of (**a**) T4 lysozyme and (**b**) neurotensin. FCA modes are sorted by fluctuation amplitude.*

both proteins FCA increased the anharmonicity of the 10 most anharmonic modes on average by more than one order of magnitude.

This result suggests the high anharmonicity as a possible reason for the improved resolution of conformational states obtained by FCA as documented in the two previous sections. Indeed, the two PCA modes that were unable to resolve the conformational states of NT sufficiently well [cf. Fig. 8(a)], show both a lower anharmonicity than their corresponding FCA modes [dashed arrows, Fig. 13(b)]. The other two labeled FCA modes [solid arrows, Fig. 13(b)] improved the resolution of conformational substates even further [cf. Fig. 8(b)], in agreement with their high anharmonicity.

Note that for T4L the fluctuation amplitude of modes, which is color coded in Figure 13(a), is uncorrelated to both, collectivity and anharmonicity. High amplitude modes (red) are seen to occur everywhere in the plot, even for those FCA modes with low collectivity and high anharmonicity, which describe the largely irrelevant swiveling motion of the terminals. For NT, fluctuation amplitude correlated with collectivity but not with anharmonicity. Hence, a selection of functionally relevant modes based purely on amplitude is expected to be less informative than one based on a combination of high collectivity and high anharmonicity.

Remaining correlations between pairs of modes

As noted earlier, FCA differs from PCA in its criterion to select modes that are the least coupled, whereas PCA identifies modes based on maximal motional amplitude. In the previous section, we have seen that in spite of these different objectives both methods yield coordinates that show remarkably similar amplitudes of atomic displacement (cf. Fig. 11). In this section we want to analyze to what extent the objective of FCA to minimize mutual information actually reduces the coupling between modes compared to PCA.

To this aim we quantified the correlation between pairs of modes for both approaches. Figure 14(a) shows that only the first 10 PCA modes of T4L exhibit large mutual correlations ($r_{MI} > 0.2$). FCA reduces these correlations to a certain degree, although the reduction may be less pronounced than expected [cf. Fig. 14(b)]. However, the small correlations, which occurred sporadically between

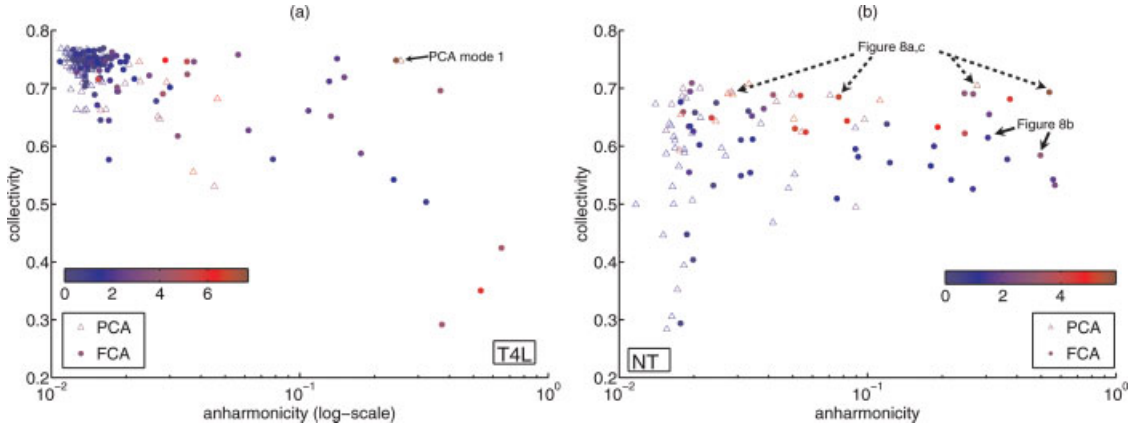

Figure 13

*The collectivity of PCA and FCA modes is plotted against their anharmonicity. The color gradient from blue to red is in accordance to an increasing fluctuation amplitude of the respective modes (quantified as $\log(V_i/V_{min})$, where $V_i = \langle c_i^2 \rangle$) (**a**) T4 lysozyme. (**b**) neurotensin; the arrows mark those modes which have been used above to determine the free energy surface in the respective figures.*

DOI 10.1002/prot

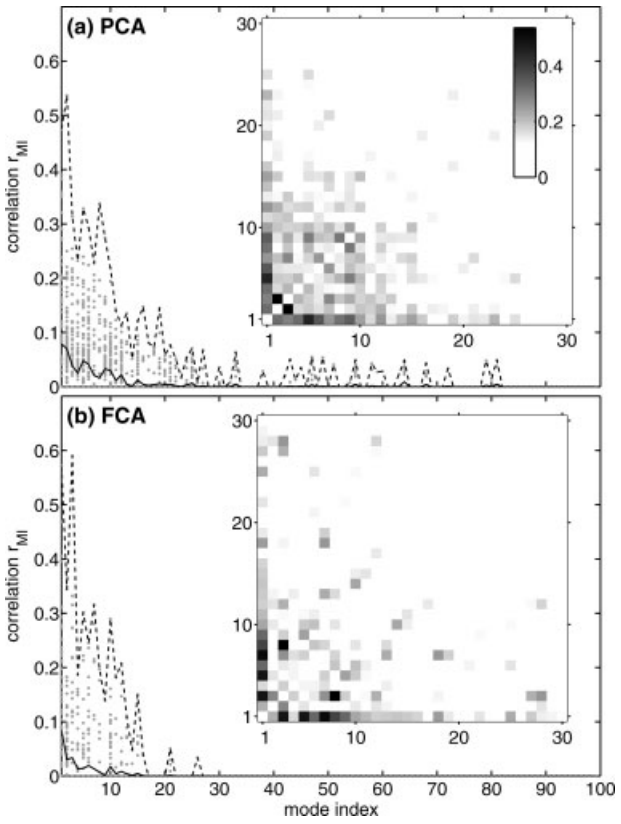

Figure 14

*Mutual correlation between pairs of FCA and PCA modes of T4 lysozyme. The plots show correlations between pairs $\mathbf{x}_i,\mathbf{x}_j$ quantified by the generalized correlation coefficient r_{MI} $[\mathbf{x}_i,\mathbf{x}_j]$ for PCA (**a**) and FCA (**b**). For every mode x_i (horiz. axis) the correlations with higher indexed modes r_{MI} $[\mathbf{x}_i,\mathbf{x}_j]_{j>i}$ were plotted (gray dots) together with the respective average (solid line) and the maximum correlation with a higher mode (dashed line). The inset shows, gray-scale coded, the mutual correlations between the first 30 modes.*

higher indexed PCA modes, were completely removed by FCA.

For NT the situation differs in two aspects [cf. Fig. 15(a,b)]. First, all PCA modes of NT showed significant mutual correlations. Second, correlations for both, high and low indexed modes, were drastically reduced by applying FCA, as indicated by the much lower average pair correlation for all FCA modes (solid line). The maximal pair correlations, however, (dashed lines) remain high and even increased in some instances.

As seen in the inset in Figure 15(b), remaining correlations constituted small clusters of coupled modes. Thus, FCA successfully identifies uncoupled motions in NT, but these are described by multiple linear FCA modes. This finding was expected, since, for example, a rotational motion of a side-chain requires more than one linear mode for its description.

Why did FCA not achieve a similar separation into uncoupled (multidimensional) modes for T4L? As a possible explanation, we suggest that the 117 ns MD simulation of T4L provides less complete sampling of configuration space compared to the trajectory of the much smaller NT. Thus, some modes of conformational motion of T4L were excited only once because of the short simulation time. For these modes "misleading" correlations are likely to be detected, because any coincidental excitation of two different modes yields a correlation of the respective modes in the generated MD ensemble. For longer trajectories with multiple transitions along these modes, these "misleading" correlations would vanish. In particular, T4L underwent a slow opening motion of its two domains described by FCA mode 1 [FCA mode 6 in Fig. 4(a)]. During one half of the simulation, T4L was closed and opened during the other half. Accordingly, all motions which occurred only once created "misleading" correlations with mode 1, which is also reflected by the high number of strong correlations of mode 1 to others, as seen in the inset of Figure 14(a,b).

Convergence of FCA

With respect to convergence, we would expect FCA to perform similarly as PCA. In particular, the notorious

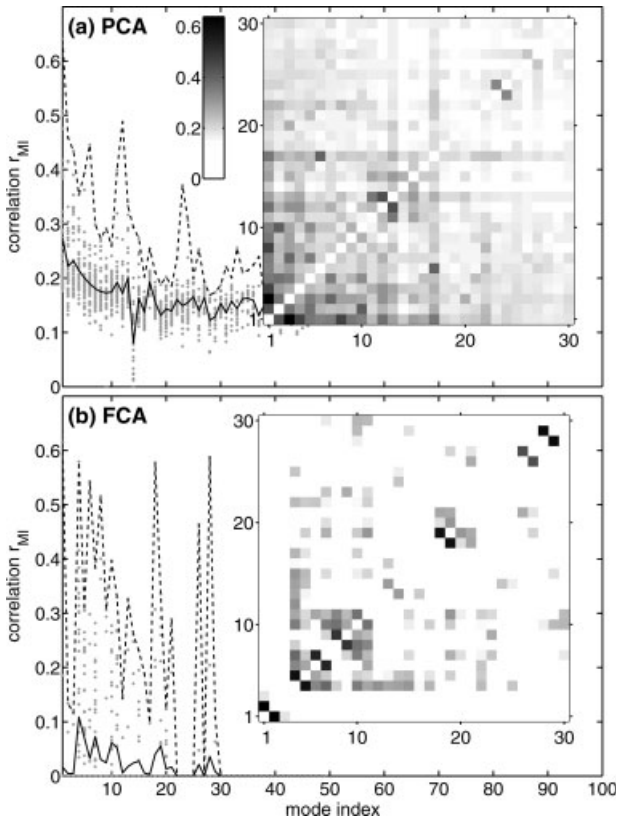

Figure 15

Correlation between pairs of FCA and PCA modes of neurotensin. The plots show correlations between pairs $\mathbf{x}_i,\mathbf{x}_j$. The inset shows the mutual correlations between the first 30 pairs of modes. For details see caption Figure 14.

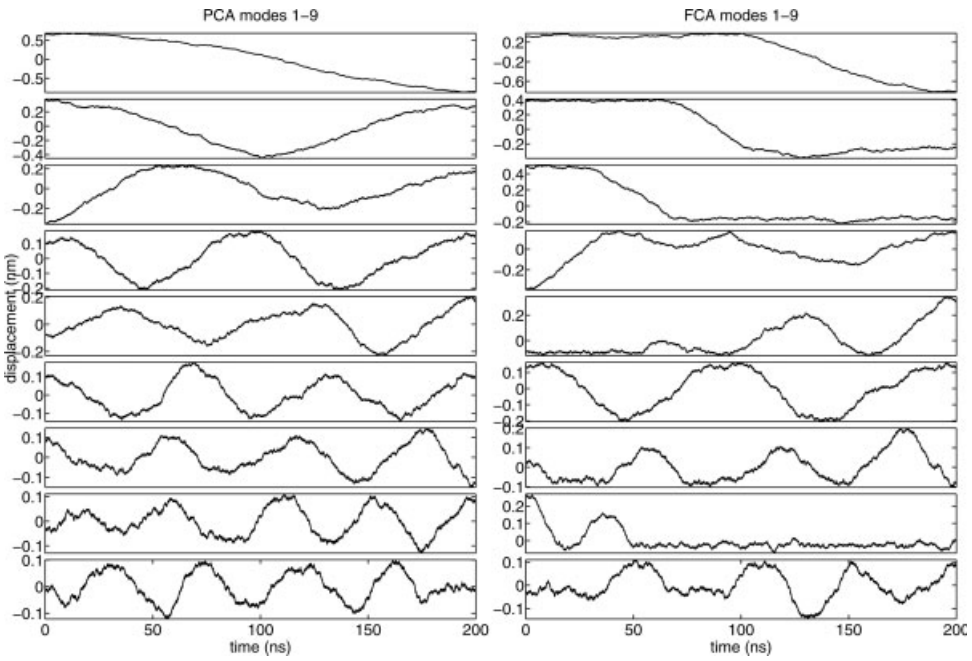

Figure 16

Projections of a 120 dimensional random walk to large amplitude PCA and FCA modes.

sampling problem in MD simulations of macromolecules will likely be reflected in similarly slow convergence of both PCA and FCA modes. The remarkable and at first sight quite surprising effect of an insufficiently sampled protein dynamics on its principal components was illustrated by Hess,[50,51] who showed that projections to principal components obtained from short MD trajectories, as for example, found in Ref. 9, are very similar to principal components of a random walk. In particular, the projections to the first PCA modes of a random walk show sine and cosine shaped curves of large amplitude,[50,51] as can be seen in Figure 16(a). This result enables one to identify those artificial large amplitude "features" in projections onto principal components that stem from incomplete sampling rather than from functionally relevant transitions between distinct conformational states.

Following these lines, we also applied FCA to a random diffusion [Fig. 16(b)]. The shown projections exhibit large amplitudes and slow transitions, as seen previously for PCA modes. However, in contrast to FCA modes of the T4L ensemble (cf. Fig. 4) the random walk FCA modes display more gradual transitions than the T4L FCA modes, suggesting that—apart from FCA mode 6—all FCA modes of T4L displayed in Figure 4 did converge to a sufficient extent. On the contrary, such a clear distinction between projections of random diffusion and the T4L ensemble cannot be established for PCA (cf. Fig. 4).

CONCLUSIONS

With FCA we have developed a new approach to extract a dimensionally reduced description of functionally relevant macromolecular motions from configurational ensembles. FCA minimizes the coupling, that is correlation, between the coordinates. In this way it differs from the well-established and widely used PCA, which maximizes the fluctuation amplitude along the coordinates.

Our comparative study of the two methods, PCA and FCA, characterized and exemplified the new method for two systems, T4 lysozyme and NT, and showed pronounced differences.

PCA on the one hand, seeking large amplitude modes, often does not identify modes that are aligned with the direction of conformational transitions. Consequently, conformational substates are not well resolved, and also more PCA modes are needed to describe a free energy surface that is consistent with the actually observed trajectory. In particular, two PCA modes did not suffice to describe the conformational motion of NT, because two otherwise separated conformational substates overlapped

DOI 10.1002/prot

in the projections to the first two PCA modes. Because of the overlap a misleadingly low free energy channel between the two conformational states emerged in a region where no transitions were seen. To provide a consistent free energy surface that resolves the barrier between distinct states a larger number of PCA modes had to be used.

By construction, and from our sample studies, FCA yields collective coordinates that are adapted to the conformational dynamics of a protein. In particular, FCA modes are typically aligned along the actual pathways of conformational transitions and thus yield an improved resolution of conformational subspaces. For both test systems, FCA modes were found to be significantly more anharmonic than their PCA counterparts. This finding also corroborates and explains the increased resolution of conformational substates by FCA. Moreover, the transition regions of free energy surfaces spanned by selected modes were found to be fully consistent with the observed transitional dynamics. This strongly suggests that FCA is a valuable alternative to PCA to yield a dimension reduced description of conformational dynamics, for example, with a Generalized Langevin framework[29] to conformational transitions.

Despite our main interest in collective motions, it is an advantageous feature of FCA to also isolate local motions, such as the swiveling motion of the terminal residues of T4 lysozyme. For PCA, this large amplitude motion was distributed over many modes, thus obscuring other more relevant motions.

Usually, the collective modes with largest amplitudes are used to analyze conformational motions of proteins.[9] However, for T4 lysozyme, this criterion selected also large amplitude local motions such as the above swiveling motions, which are unlikely to play an important functional role in substrate binding. In such cases, where amplitude does not point to functionally relevant modes, and following the suggestion by Amadei *et al.*[52] that functional motion often implies anharmonicity, one could select modes by their anharmonicity rather than their amplitude. However, the FCA modes that described the irrelevant swiveling motion were not only of large amplitude but also highly anharmonic. Yet, they showed a very low collectivity, such that we suggest to rank FCA modes by a combination of the two properties, anharmonicity and collectivity.

We found it helpful to visualize the matrix of mutual correlations of FCA modes as shown in the inset of Figure 15b. On the basis of these pair correlations, a scheme was proposed that selects those pairs of modes that are particularly suitable to visualize the structure of the essential subspace (cf. Fig. 5). Moreover, the analysis of pair correlations revealed that for neurotensin FCA successfully separated the dynamics into several uncoupled motions, whereas this was achieved only to a limited extent for the larger T4 lysozyme, where in particular the slow mode 1 remained strongly correlated to all other modes. Close analysis suggested that this partial failure is not inherent in FCA, but rather must be attributed to the insufficient sampling of the slow opening motion of the two T4L domains.

As shown, FCA represents a valuable alternative to PCA. The two methods, FCA and PCA, use different optimization criteria to define internal coordinates, and thus the choice to select the one method over the other will depend on the questions asked. We have shown that FCA improves the resolution and separation of conformational states, whereas many other useful features of PCA modes are preserved. For instance, conformational entropies of proteins are usually computed with PCA from MD simulation trajectories.[12,13] An upper bound of this information-entropy is estimated by assuming independent harmonic oscillators for every PCA mode. A significantly improved upper bound is expected from maximally uncoupled modes extracted by FCA. Further improvements are possible from combining FCA modes which have a high remaining pair correlation into several low-dimensional subspaces, and to estimate the information-entropy of these subspaces directly with nearest neighbor approaches.

Further work needs to address convergence issues. Firstly, the convergence and robustness of the minimization of mutual information needs to be examined. In particular, it needs to be addressed if the global minimum of mutual information is always actually found, and whether similar FCA modes are extracted from slightly perturbed MD ensembles. Secondly, the slow convergence of correlations in the configurational ensemble due to the sampling problem of MD[53] needs to be analyzed. The finding that FCA—just like PCA—yields highly anharmonic modes for a random diffusion[51] suggest that in this respect the convergence properties are very similar. Nonetheless, the application of FCA to a random walk indicated that the "foot-print" of an unconverged mode is more distinctly identified in projections to FCA modes than for PCA modes.

Finally, it has been suggested to use nonlinear coordinates for PCA.[41,43] Similarly, it will be rewarding to combine nonlinear coordinates with the criterion of minimizing mutual information.

ACKNOWLEDGMENTS

We thank Bert L. de Groot for providing the trajectories of neurotensin and T4 Lysozyme, and Ira Tremmel and Lars V. Schäfer for carefully reading the manuscript. This work was supported by the Volkswagen Foundation, grant I/80436.

REFERENCES

1. Berendsen HJC, Hayward S. Collective protein dynamics in relation to function. Curr Opin Struct Biol 2000;10:165–169.

2. Norberg J, Nilsson L. Advances in biomolecular simulations: methodology and recent applications. Q Rev Biophys 2003;36:257–306.
3. Kitao A, Gō N. Investigating protein dynamics in collective coordinate space. Curr Opin Struct Biol 1999;9:164–169.
4. Brooks B, Karplus M. Harmonic dynamics of proteins–normal-modes and fluctuations in bovine pancreatic trypsin-inhibitor. Proc Natl Acad Sci USA 1983;80:6571–6575.
5. Gō N, Noguti T, Nishikawa T. Dynamics of a small globular protein in terms of low-frequency vibrational-modes. Proc Natl Acad Sci USA 1983;80:3696–3700.
6. Levitt M, Sander C, Stern PS. The normal-modes of a protein–native bovine pancreatic trypsin-inhibitor. Int J Quantum Chem 1983; (Suppl. 10):181–199.
7. Kitao A, Hirata F, Gō N. The effects of solvent on the conformation and the collective motions of protein–normal mode analysis and molecular-dynamics simulations of melittin in water and in vacuum. Chem Phys 1991;158:447–472.
8. Garcia AE. Large-amplitude nonlinear motions in proteins. Phys Rev Lett 1992;68:2696–2699.
9. Amadei A, Linssen ABM, Berendsen HJC. Essential dynamics of proteins. Proteins 1993;17:412–425.
10. Qian W, Bandekar J, Krimm S. Vibrational analysis of peptides, polypeptides, and proteins. 41. Vibrational analysis of crystalline tri-L-alanine. Biopolymers 1991;31:193–210.
11. Frauenfelder H, Wolynes PG. Rate theories and puzzles of hemeprotein kinetics. Science 1985;229:337–345.
12. Karplus M, Kushick JN. Method for estimating the configurational entropy of macromolecules. Macromolecules 1981;14:325–332.
13. Schlitter J. Estimation of absolute and relative entropies of macromolecules using the covariance matrix. Chem Phys Lett 1993;215: 617–621.
14. Duda RO, Hart PE, Stork DG. Pattern classification. New York: Wiley; 2001.
15. Cover TM, Thomas JA. Elements of information theory. New York: Wiley; 1991.
16. Cardoso JF. Blind signal separation: statistical principles. Proc IEEE 1998;86:2009–2025.
17. Fraser AM, Swinney HL. Independent coordinates for strange attractors from mutual information. Phys Rev A 1986;33:1134–1140.
18. Comon P. Independent component analysis, a new concept. Signal Process 1994;36:287–314.
19. Hyvärinen A. Fast and robust fixed-point algorithms for independent component analysis. IEEE Trans Neural Networks 1999;10:626–634.
20. Stogbauer H, Kraskov A, Astakhov SA, Grassberger P. Least-dependent-component analysis based on mutual information. Phys Rev E 2004;70:066123.
21. Almeida LB. MISEP–linear and nonlinear ICA based on mutual information. J Mach Learn Res 2004;4:1297–1318.
22. Lange OF, Grubmüller H. Generalized correlation for biomolecular dynamics. Proteins 2006;62:1053–1061.
23. Hyvärinen A, Karhunen J, Oja E. Independent component analysis. Wiley, New York; 2001.
24. Forsythe GE, Malcolm MA, moler CB. Computer methods for mathematical computations. New York: Prentice-Hall; 1976.
25. Learned-Miller EG, Fisher JW. ICA using spacings estimates of entropy. J Mach Learn Res 2004;4:1271–1295.
26. Moon YI, Rajagopalan B, Lall U. Estimation of mutual information using kernel density estimators. Phys Rev E 1995;52:2318–2321.
27. Wand MP. Data-based choice of histogram bin width. Am Stat 1997;51:59–64.
28. Hayward S, Kitao A, Gō N. Harmonicity and anharmonicity in protein dynamics–a normal-mode analysis and principal component analysis. Proteins 1995;23:177–186.
29. Lange OF, Grubmüller H. Collective Langevin dynamics of conformational motions in proteins. J Chem Phys 2006;124:214903.
30. Luca S, White JF, Sohal AK, Filippov DV, vanBoom JH, Grisshammer R, Baldus M. The conformation of neurotensin bound to its G protein-coupled receptor. Proc Natl Acad Sci USA 2003;100:10706–10711.
31. Lindahl E, Hess B, Van der Spoel D. GROMACS 3.0: a package for molecular simulation and trajectory analysis. J Mol Model 2001;7: 306–317.
32. Jorgensen WL, Maxwell DS, Tirado-Rives J. Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. J Am Chem Soc 1996;118:11225–11236.
33. Hess B, Bekker H, Berendsen HJC, Fraaije JGEM. LINCS: a linear constraint solver for molecular simulations. J Comput Chem 1997; 18:1463–1472.
34. Miyamoto S, Kollman PA. SETTLE: an analytical version of the SHAKE and RATTLE algorithms for rigid water models. J Comput Chem 1992;13:952–962.
35. Darden T, York D, Pedersen L. Particle mesh Ewald: an N·log(N) method for Ewald sums in large systems. J Chem Phys 1993;98: 10089–10092.
36. Essmann U, Perera L, Berkowitz ML, Darden T, Lee H, Pedersen LG. A smooth particle mesh Ewald method. J Chem Phys 1995;103: 8577–8593.
37. Berendsen HJC, Postma JPM, DiNola A, Haak JR. Molecular dynamics with coupling to an external bath. J Chem Phys 1984;81: 3684–3690.
38. Kraskov A, Stogbauer H, Grassberger P. Estimating mutual information. Phys Rev E 2004;69:066138.
39. Shwartz S, Zibulevsky M, Schechner YY. Fast kernel entropy estimation and optimization. Signal Process 2005;85:1045–1058.
40. Lange OF. g_fca (c) 2006; http://www.mpibpc.mpg.de/groups/grubmueller/olange/fca.html.
41. Nguyen PH. Complexity of free energy landscapes of peptides revealed by nonlinear principal component analysis. Proteins Struct Funct Bioinfo 2006;65:898–913.
42. Kramer MA. Nonlinear principal component analysis using autoassociative neural networks. Aiche J 1991;37:233–243; 32.
43. Tenenbaum JB, deSilva V, Langford JC. A global geometric framework for nonlinear dimensionality reduction. Science 2000;290:2319–2323.
44. Hyvärinen A, Oja E. Independent component analysis: algorithms and applications. Neural Networks 2000;13:411–430.
45. Kuroki R, Weaver LH, Mathhews BW. A covalent enzyme-substrate intermediate with saccharide distortion in a mutant T4 lysozyme. Science 1993;262:2030.
46. Faber HR, Matthews BW. A mutant T4 lysozyme displays five different crystal conformations. Nature 1990;348:263–266.
47. Lu HP. Single-molecule spectroscopy studies of conformational change dynamics in enzymatic reactions. Curr Pharm Biotechnol 2004;5:261–269.
48. Mchaourab HS, Oh KJ, Fang CJ, Hubell WL. Conformation of T4 lysozyme in solution. Hinge-bending motion and the substrate-induced conformational transition studied by site-directed spin labeling. Biochemistry 1997;36:307–316.
49. de Groot BL, Hayward S, van Aalten DMF, Amadei A, Berendsen HJC. Domain motions in bacteriophage T4 lysozyme: A comparison between molecular dynamics and crystallographic data. Proteins 1998;31:116–127.
50. Hess B. Similarities between principal components of protein dynamics and random diffusion. Phys Rev E 2000;62:8438–8448.
51. Hess B. Convergence of sampling in protein simulations. Phys Rev E 2002;65:031910.
52. Amadei A, de Groot BL, Ceruso MA, Paci M, Nola AD, Berendsen HJC. A kinetic model for the internal motions of proteins: Diffusion between multiple harmonic wells. Proteins: Struct Funct Genet 1999;35:283–292.
53. Clarage JB, Romo T, Andrews BK, Pettitt BM, Phillips GN. A sampling problem in molecular-dynamics simulations of macromolecules. Proc Natl Acad Sci USA 1995;92:3288–3292.

 DOI 10.1002/prot