OPEN ACCESS Freely available online

Detection of Functional Modes in Protein Dynamics

Jochen S. Hub, Bert L. de Groot*

Computational Biomolecular Dynamics Group, Max-Planck-Institute for Biophysical Chemistry, Göttingen, Germany

Abstract

Proteins frequently accomplish their biological function by collective atomic motions. Yet the identification of collective motions related to a specific protein function from, e.g., a molecular dynamics trajectory is often non-trivial. Here, we propose a novel technique termed "functional mode analysis" that aims to detect the collective motion that is directly related to a particular protein function. Based on an ensemble of structures, together with an arbitrary "functional quantity" that quantifies the functional state of the protein, the technique detects the collective motion that is maximally correlated to the functional quantity. The functional quantity could, e.g., correspond to a geometric, electrostatic, or chemical observable, or any other variable that is relevant to the function of the protein. In addition, the motion that displays the largest likelihood to induce a substantial change in the functional quantity is estimated from the given protein ensemble. Two different correlation measures are applied: first, the Pearson correlation coefficient that measures linear correlation only; and second, the mutual information that can assess any kind of interdependence. Detecting the maximally correlated motion allows one to derive a model for the functional state in terms of a single collective coordinate. The new approach is illustrated using a number of biomolecules, including a polyalanine-helix, T4 lysozyme, Trp-cage, and leucine-binding protein.

Citation: Hub JS, de Groot BL (2009) Detection of Functional Modes in Protein Dynamics. PLoS Comput Biol 5(8): e1000480. doi:10.1371/journal.pcbi.1000480

Editor: Ruth Nussinov, National Cancer Institute, United States of America and Tel Aviv University, Israel

Received April 20, 2009; **Accepted** July 21, 2009; **Published** August 28, 2009

Funding: This study was supported by the Max-Planck-Society (http://www.mpg.de/english/portal/index.html). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: bgroot@gwdg.de

Introduction

Collective motions are essential for biological functions in proteins [1]. They are involved in numerous biological processes including enzyme catalysis, channel gating, allosteric interactions, signal transduction, and recognition dynamics. The observed motions are as diverse as hinge, shear, or rotational motions of entire subunits, opening motions of molecular lids, loop motions, partial unfolding, or subtle rearrangements of amino acid side chains [2]. Understanding the functional mechanisms of such proteins requires both to identify the protein's collective motions and to relate the observed motions to the protein's biological function.

Diverse experimental methods have been applied to elucidate collective motions including nuclear magnetic resonance (NMR) [3,4], X-ray crystallography [5], as well as single-molecule fluorescence [6] or electron-transfer measurements [7]. Complementary to experiments, molecular dynamics (MD) simulations are a widely used techniques to investigate collective motions in proteins [8]. A state-of-the-art approach to elucidate collective motions from the protein dynamics is principal component analysis (PCA) [9–12]. PCA is commonly used to extract the collective motions with the largest contribution to the variance of the atomic fluctuations. Alternatively, normal mode analysis (NMA) has been extensively used to identify low-frequency collective modes [13–16]. Such modes are expected to correspond to large atomic displacements and are therefore assumed to be important to protein function. In addition, elastic network models are an established approach to assess motions intrinsic to the protein structure [17].

Established methods such as PCA and NMA elucidate large-scale and low-frequency modes, respectively, but do not necessarily yield collective motions directly related to protein function. Here, we propose a novel analysis technique termed 'functional mode analysis' (FMA) that aims to elucidate collective motions directly related to a specific protein function. As input, the technique requires a set of protein structures, together with a 'functional quantity' that can be expressed as a single number for each input structure. The structures typically derive from an MD simulation, but a (large) set of X-ray or NMR structures is equally well suited. The chosen 'functional quantity' can be quite general and could correspond to some geometric, electrostatic, or chemical observable, or any other variable that might be relevant to the function of the protein. Typical examples for the functional quantity could include the openness of a channel, active site geometry, or cleft solvent accessibility. Given that input, the technique seeks the collective protein motion that is maximally related to the functional quantity. In other words, the technique aims to explain variations in the functional quantity in terms of collective motions.

When relating a functional quantity (from now on termed f) to collective motions, two quite different motions might be of interest. First, the motion that displays the largest correlation to f. This motion is unaffected by the energy landscape of the protein, and it will be referred to as 'maximally correlated motion' (MCM). It is particularly interesting for quantities f of which the dependence on the protein structure is complex and therefore unclear. An example for such a complex quantity would be the R-value in X-ray refinement. Second, the physical motion that actually accomplishes substantial deviations in f, in accordance with the protein's energy landscape, is frequently of interest. Because many different motions might affect f, we use the input structure ensemble to estimate the most probable collective motion that

Author Summary

Proteins are flexible nanomachines that frequently accomplish their biological function by collective atomic motions. Such motions may be characterized by hinge, shear, or rotational motions of entire protein domains, loop movements, or subtle rearrangements of amino acid side chains. In many cases it is far from obvious how collective motions are related to a particular biological task. Therefore, we propose a novel technique termed "functional mode analysis" that, based on an ensemble of structures, aims to detect a collective motion that is directly related to a particular protein function. From the given set of protein structures, together with a "functional quantity", the technique seeks the collective motion that is maximally correlated to the functional quantity. The chosen functional quantity can be quite general; typical examples could include the openness of a channel, active site geometry, or cleft solvent accessibility. Because the proposed framework is highly general, we expect the approach to be useful to a wide range of applications. To illustrate the new technique, we apply functional mode analysis to molecular dynamics trajectories of a polyalanine-helix, bacteriophage T4 lysozyme, Trp-cage, and Leucine-binding protein.

accomplishes a substantial change in f. That motion will be referred to as ensemble-weighted MCM (ewMCM). Depending on the question addressed, the MCM or the ewMCM (or both) can provide insight into the relation between function and motion. Therefore, both motions are considered by the proposed framework.

This paper is organized as follows. First, we describe the analysis technique. Subsequently, four examples for FMA are presented, applying the approach to a polyalanine helix, T4 lysozyme, Trp-cage, and leucine-binding protein. The examples illustrate the use of FMA in detecting functionally relevant collective motions.

Methods

Theory and Concepts

Let us consider the simulation trajectory $\mathbf{x}(t)\in\mathbb{R}^{3N}$ of the protein atoms or of a subset of the protein atoms such as the backbone or the heavy atoms. $\mathbf{x}(t)$ denotes the $3N$ cartesian coordinates of N atoms. The coordinates are known at N_t times, i.e., $t\in\{t_1,\ldots,t_{N_t}\}$. For each time t, an arbitrary scalar functional quantity $f(t)$ is given which can be computed from the protein coordinates and/or velocities. Note that for the following presentation of FMA, the time t is only an index to label the input structures $\mathbf{x}(t)$ and should not imply that the structures must correspond to a time series. Instead, the structures $\mathbf{x}(t)$ may equally well derive from, e.g., Monte Carlo sampling or from a large ensemble of experimental structures.

Maximally correlated motion (MCM). We seek a normalized collective vector $\mathbf{a}\in\mathbb{R}^{3N}$ of protein atoms such that the motion along $\mathbf{a}$ is maximally correlated to the change in the functional quantity $f(t)$. Therefore, the motion along $\mathbf{a}$ is referred to as 'maximally correlated motion' (MCM). The MCM as a function of time t is given by the projection

$$p_a(t)=[\mathbf{x}(t)-\langle\mathbf{x}\rangle]\cdot\mathbf{a}, \tag{1}$$

where $\langle\cdots\rangle$ denotes the average over all times t.

In the present study, two measures are applied to quantify the correlation between f and p_a. First, Pearson's correlation coefficient defined by

$$R=\frac{\mathrm{cov}(f,p_a)}{\sigma_f\sigma_a}, \tag{2}$$

where $\mathrm{cov}(f,p_a)$ denotes the covariance between $f(t)$ and $p_a(t)$, and σ_f and σ_a denote the standard deviations of $f(t)$ and $p_a(t)$, respectively. The Pearson coefficient measures only linear correlation. Second, the mutual information (MI) between f and p_a given by [18]

$$I(f,p_a)=\int\int P(f',p'_a)\log\left(\frac{P(f',p'_a)}{P_1(f')P_2(p'_a)}\right)\mathrm{d}f'\mathrm{d}p'_a. \tag{3}$$

Here, $P(f',p'_a)$ denotes the joint probability distribution of f and p_a, and $P_1(f')$ and $P_2(p'_a)$ denote the marginal probability distributions of f and p_a, respectively. The MI measures any kind of interdependence between f and p_a, including non-linear and higher order correlation. Note that if (and only if) f and p_a are independent, $P(f',p'_a)=P_1(f')P_2(p'_a)$ holds, the logarithm in eq. (3) vanishes, and the MI equals to zero. Hence, the MI can be interpreted as the probability weighted deviation from the case of f and p_a being independent.

Reduction of dimensionality. Before optimizing $\mathbf{a}$ (via maximization of R or of the MI), a reduction of the dimensionality of the optimization problem is frequently required. Even when restricting the analysis to a subset of the protein atoms (such as the backbone), the long autocorrelations in protein dynamics may otherwise lead to an overfitted collective vector $\mathbf{a}$. A common procedure to reduce the dimensionality of protein dynamics is principal component analysis (PCA) [10]. PCA allows one to determine a small set of collective vectors with the largest contribution to the mean square fluctuations (MSF) of the atomic coordinates.

For convenience and to clarify the nomenclature we briefly sketch the PCA in the following. Given the $3N$ cartesian atomic coordinates $x_i(t)$ $(i=1,\ldots,3N)$, the elements of the covariance matrix C of the atomic positions are given by

$$C_{ij}=\langle(x_i-\langle x_i\rangle)(x_j-\langle x_j\rangle)\rangle. \tag{4}$$

Before computing C, translation and rotation of the entire biomolecule is removed by superimposing the simulation trajectory onto a reference structure. Diagonalization of C yields a set of $3N$ orthonormal eigenvectors $\mathbf{e}_i$ with corresponding eigenvalues σ_i^2. The eigenvectors are typically ordered according to descending eigenvalues and referred to as PCA vectors. The projection $p_i(t)=[\mathbf{x}(t)-\langle\mathbf{x}\rangle]\cdot\mathbf{e}_i$ is called i^{th} principal component (PC) and quantifies the position of the protein along the i^{th} PCA vector.

The MSF of the atoms can be decomposed into contributions from different principal components, $\langle(\mathbf{x}-\langle\mathbf{x}\rangle)^2\rangle=\sum_{i=1}^{3N}\mathrm{var}(p_i)=\sum_{i=1}^{3N}\sigma_i^2$, where $\mathrm{var}(\cdots)$ denotes the variance. In protein simulations the first 10–20 PCs (with the largest eigenvalues) often account for a large fraction (80–90%) of the atomic MSF, and higher PCs describe smaller motions such as angle vibrations [10]. Hence, if large protein motions are expected to dominate changes in the $f(t)$, the first few PCA vectors are a reasonable basis set to construct $\mathbf{a}$.

We stress that PCA vectors are only one possible basis for $\mathbf{a}$. Other possibilities include normal modes or modes derived from

full correlation analysis [16,19]. For some quantities $f(t)$ angles in dihedral space or the cartesian coordinates may also provide a useful coordinate system.

Maximization of the Pearson coefficient R. Assuming that f is approximately a linear function of the PCs, the collective vector $\mathbf{a}$ can be derived by maximizing the Pearson coefficient R (eq. (2)). We construct $\mathbf{a}$ as a linear combination of the first d PCA vectors, $\mathbf{a}=\sum_{i=1}^{d}\alpha_i\mathbf{e}_i$. Here, coefficients α_i denote the coordinates of $\mathbf{a}$ with respect to the basis set $\{\mathbf{e}_i\}$.

As shown in the supporting Text S1, a maximum in the absolute value of R can be found by numerically solving the coupled linear set of equations

$$\sum_{i=1}^{d}\alpha_i\mathrm{cov}(p_i,p_\ell)=\mathrm{cov}(f,p_\ell),\ \ \ell=1,\ldots,d. \tag{5}$$

For the present study, $\mathbf{a}=(\alpha_1,\ldots,\alpha_d)$ was normalized after computation via eqs. (5). Note that for the maximization of R, the normalization of $\mathbf{a}$ is not strictly necessary because R is invariant to the norm of $\mathbf{a}$.

It is instructive to note that maximizing R provides a quantitative model for f as a function of the PCs $p_i(t)$. A model for f allows one to predict f from a given protein structure, and, in turn, propose new structures that generate a particular value of f (i.e., a specific functional state). Let

$$m_f(t)=\langle f\rangle+\sum_{i=1}^{d}\beta_i[p_i(t)-\langle p_i\rangle] \tag{6}$$

denote the model for $f(t)$. The d parameters β_i are fitted to the data $f(t)$ by minimizing the mean square deviation (MSD) between $f(t)$ and its model $m_f(t)$, i.e., $\langle[f-m_f]^2\rangle\rightarrow\min$. In multivariate regression analysis this approach is referred to as 'ordinary least square estimation' [20]. The minimum of $\langle[f-m_f]^2\rangle$ with respect to the parameters β_i is found by solving the coupled set of linear equations

$$\sum_{i=1}^{d}\beta_i\mathrm{cov}(p_i,p_\ell)=\mathrm{cov}(f,p_\ell),\ \ \ell=1,\ldots,d, \tag{7}$$

as shown in Text S1. By comparison to eqs. (5), the β_i are identical to α_i with the exception that the α_i may be scaled by an arbitrary factor without changing R. Hence, $m_f(t)$ can be rewritten in terms of p_a via $m_f(t)=\langle f\rangle+v(p_a(t)-\langle p_a\rangle)$, where v is a constant.

Maximization of the mutual information (MI). If f depends non-linearly on atomic positions, the Pearson coefficient might be an insufficient measure to detect correlation between f and p_a. In such cases, we apply the MI as correlation measure because it can detect any kind of interdependence. Optimizing the MI is computationally more demanding as compared to the optimization of R. The methodological details for the MI optimization are described in the section 'Iterative optimization of the mutual information'.

Maximizing the MI yields the collective vector p_a that, by construction, provides as much information on f as possible. Because the functional relation between p_a and f can be arbitrary, optimizing the MI does not directly provide a quantitative model for f (as in the case of optimizing R, compare previous paragraph). A natural approach to nevertheless yield a model for f is to fit a general curve $g(p_a)$ (such as a spline [21] or a polynomial) to the p_a-f data points. Given a protein structure $\mathbf{x}$, this model allows one to predict the quantity f from the structure via $f\approx g((\mathbf{x}-\langle\mathbf{x}\rangle)\cdot\mathbf{a})$.

Contributions of principal components to $f(t)$. PCA modes have frequently been shown to be important to protein function [8,10,22]. However, a functionally important motion may be spread over a number of PCA modes. To understand the protein's function, it is therefore instructive to quantify the influence of different PCA modes on the functional quantity $f(t)$, in particular if the PCA modes are related to intuitive motions such as hinge-bending or torsional modes.

Let us first consider the case of a linear model for f (eq. (6)). Using the linear model for $m_f(t)$ as an approximation for $f(t)$ (eq. (6)), the variance of $f(t)$ can be approximated by

$$\mathrm{var}(f)\approx\mathrm{var}(m_f)=\sum_{i=1}^{d}\left[\beta_i^2\mathrm{var}(p_i)+\beta_i\sum_{j\neq i}\beta_j\mathrm{cov}(p_i,p_j)\right]. \tag{8}$$

The expression in square brackets is the contribution of the i^{th} PC to the variance of $f(t)$. When using the same set of simulation frames for the PCA and for constructing the model $m_f(t)$, all $\mathrm{cov}(p_i,p_j)$ vanish for $i\neq j$ and eq. (8) simplifies to $\mathrm{var}(f)\approx\sum_{i=1}^{d}\beta_i^2\mathrm{var}(p_i)$.

In case of a non-linear dependence between f and p_a, $\mathrm{var}(f)$ cannot be decomposed into contributions from different PCs. Instead, the variance of p_a can be written as the right hand side of eq. (8), except for the β_i being substituted by the α_i. This way, fluctuations of the motion correlated to f (but not the variance in f itself) can be decomposed into contributions from different PCs.

Ensemble-weighted maximally correlated motion contributing to f. The MCM along the collective vector $\mathbf{a}$ displays the largest correlation to $f(t)$ as measured from R or from the MI. However, due to the protein's energy landscape, the motion parallel to $\mathbf{a}$ may be severely restricted. This fact is schematically illustrated in Fig. 1. Let us assume that the functional quantity f increases to the right in Fig. 1. Then, irrespective of the energy landscape (thin lines in Fig. 1A–C), the MCM is parallel to the direction of increasing f. However, if the energy landscape restricts the motion parallel to the MCM (Fig. 1B), a displacement along the MCM will actually occur through a motion which is non-parallel to the MCM, but which is in accordance to the energy landscape. Such motions would have a substantial projection on the MCM, but are not identical to the MCM.

In addition to the MCM, we therefore seek the *most probable* collective motion that accomplishes a specific displacement along the MCM. We apply the input structure ensemble to estimate the most probable motion, and refer to that motion as 'ensemble-weighted maximally correlated motion' (ewMCM). The ewMCM is shown as black arrows in Fig. 1 for three different hypothetical energy landscapes. Note that the ewMCM is pointing in the direction that accomplishes a displacement along the MCM with a motion of smallest energy increase (i.e., with the largest probability). If the energy landscape does not favor any direction (Fig. 1A), the ewMCM is parallel to the MCM. In contrast, if the energy landscape highly favors motions in one direction over motions in another direction (Fig. 1B), the ewMCM may strongly deviate from the MCM. An intermediate situation is shown in Fig. 1C. It should be emphasized that the ewMCM is the collective motion *in the given input ensemble* which accomplishes the displacement along the MCM. In case of limited sampling, a second input ensemble may accomplish a displacement along the MCM through a different ewMCM, rendering the ewMCM highly dependent on the input

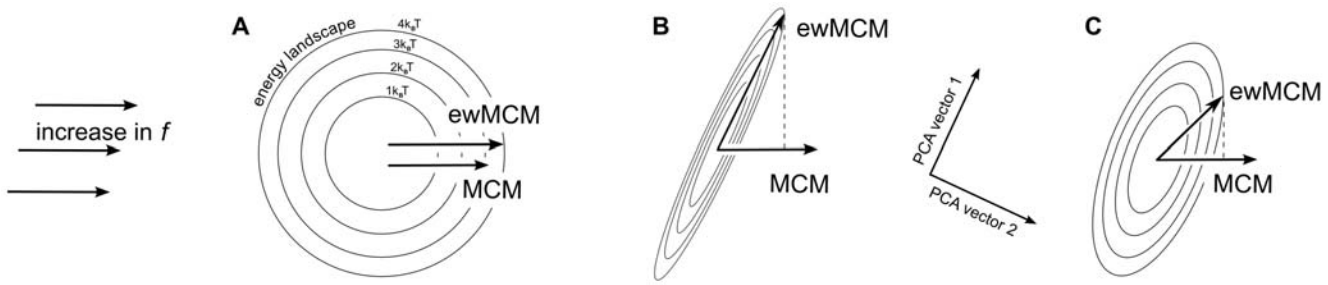

Figure 1. On the difference between the maximally correlated motion (MCM) and the ensemble-weighted MCM (ewMCM) contributing to the functional quantity *f*. (A–C) Irrespective of the energy landscape (thin lines), the MCM (along **a**) is parallel to the direction with increasing f (here, to the right). In contrast, the ewMCM is highly dependent on the energy landscape. (A) If no direction is favored by the energy landscape, the ewMCM is parallel to the MCM. (B) If one direction (PCA vector 1) is highly favored over another direction (PCA vector 2), a displacement along the MCM is mainly accomplished through a motion along the PCA vector 1. Therefore, the ewMCM is nearly parallel to PCA vector 1 in this case. (C) An intermediate situation between the extreme cases (A) and (B). In that case, both PCA vectors 1 and 2 contribute to the ewMCM.
doi:10.1371/journal.pcbi.1000480.g001

ensemble. This characteristic of the ewMCM motivates the term 'ensemble-weighted'.

Let us assume that the MCM $\mathbf{a}=\sum_{i=1}^{d}\alpha_i\mathbf{e}_i$ has been optimized such that $p_a(t)=\sum_{i=1}^{d}\alpha_i p_i(t)$ is maximally correlated to $f(t)$, as measured from R or from the MI (see above). Thus, the set of α_i are fixed in the following, and p_a quantifies the functional state of the protein. In the input ensemble, the collective variable p_a varies between its minimum $p_a^{\min}=\min(p_a(t),t\in\{t_1,\ldots,t_{N_t}\})$ and its maximum $p_a^{\max}=\max(p_a(t),t\in\{t_1,\ldots,t_{N_t}\})$. To define the ewMCM, we choose an arbitrary but fixed value for p_a, denoted $p_a^*\in[p_a^{\min},p_a^{\max}]$, between its extremes $p_a^{\min}$ and $p_a^{\max}$. As ewMCM we consider the *most probable* collective displacement $\mathbf{v}^*$ (from the average structure $\langle\mathbf{x}\rangle$) that generates the functional state p_a^*,

$$P(\mathbf{v}^*)\rightarrow\max,\quad \mathbf{v}^*\cdot\mathbf{a}=p_a^*. \tag{9}$$

Here, $P(\mathbf{v}^*)$ denotes the probability of the collective atomic displacement $\mathbf{v}^*$. In the following we restrict the ewMCM $\mathbf{v}^*$ to the subspace spanned by the first d PCA vectors $\{\mathbf{e}_i\}$. Then, the ewMCM can be expressed as $\mathbf{v}^*=\sum_{i=1}^{d}p_i^*\mathbf{e}_i$ and the condition (9) can be rewritten as

$$P(p_1^*,\ldots,p_d^*)\rightarrow\max,\quad \sum_{i=1}^{d}\alpha_i p_i^*=p_a^*, \tag{10}$$

where $P(p_1^*,\ldots,p_d^*)$ denotes the probability for a particular set of PCs $p_1^*,\ldots,p_d^*$.

The p_i^* were estimated as follows. First, to simplify the nomenclature, let assume the mutual covariances between the PCs to equal zero. Then, $P(p_1,\ldots,p_d)$ can be approximated via

$$P(p_1,\ldots,p_d)\approx\prod_{i=1}^{d}P_i(p_i)\approx N_p^{-1}\exp\left(-1/2\sum_{i=1}^{d}p_i^2/\sigma_i^2\right), \tag{11}$$

where P_i denotes the marginal probability distribution of the i^{th} PC, and N_p is a normalization constant. Here, the p_i were assumed to be mutually independent and normally distributed. (If the PCs were constructed from a different set of frames than the frames used for the FMA, the covariances between different PCs may not vanish, rendering the assumption $P(p_1,\ldots,p_d)\approx\prod_{i=1}^{d}P_i(p_i)$ a poor approximation. In that case we switch to new coordinates $(q_1,\ldots,q_d)=\mathbf{K}(p_1,\ldots,p_d)$ with zero mutual covariances. Here, the transformation matrix $\mathbf{K}$ is computed from a PCA on the $p_1,\ldots,p_d$.) Using the approximation in eq. (11), the maximization of $P(p_1^*,\ldots,p_d^*)$ with the constraint $\sum_{i=1}^{d}\alpha_i p_i^*=p_a^*$ is straightforward using Lagrange multipliers. The calculation yields

$$p_i^*=\frac{\alpha_i\sigma_i^2}{\sum_{j=1}^{d}(\alpha_j\sigma_j)^2}\,p_a^*. \tag{12}$$

Note that the components α_i of the MCM are weighted by the variance σ_i^2 of p_i of the given ensemble, further justifying the term 'ensemble-weighted'. To visualize the ewMCM for the given ensemble, successively increasing values for p_a^* can be chosen between, e.g., $p_a^{\min}$ and $p_a^{\max}$. For each value p_a^*, eq. (12) provides a set of PCs p_i^* and, hence, a structure $\mathbf{x}=\langle\mathbf{x}\rangle+\sum_{i=1}^{d}p_i^*\mathbf{e}_i$. The structures can be depicted in common molecular visualization software.

Cross-Validation

MD simulations of proteins can be subject to long autocorrelations. The maximization of R or MI can lead to overfitting if too many free parameters α_i are used in the optimization. It is therefore essential to cross-validate the derived model for $f(t)$ with an independent set of simulation frames. A convenient approach to cross-validate the optimization is to divide the simulation into frames for model building and for cross-validation. Accordingly, R or MI is optimized applying the model building set only, yielding a correlation R_m between data and model. Subsequently, the derived model is validated by predicting $f(t)$ from the derived model using the cross-validation set only, yielding a correlation R_c. Note that in the present context the term 'predict' should *not* imply any prediction into the future. Instead, the exact (or true) $f(t)$ as computed from all atomic coordinates is compared with $f(t)$ as computed from the model, making only use of the functional collective coordinate $p_a(t)$. Hence, we apply the term 'predict' as common in, e.g., the pattern recognition literature [23]. Using this approach, overfitting is indicated by a substantially smaller R_c as compared to R_m.

How many basis vectors d should be used to construct **a**? The optimal number d will highly depend on the simulation system and the observable $f(t)$. The reasonable choice for d can be identified by plotting R_m and R_c as a function of d. As long as no overfitting occurs, R_c increases with d, indicating an improvement of the

model. As soon as R_c decreases with d or becomes substantially smaller than R_m, the model is overfitted.

Iterative Optimization of the Mutual Information

The collective vector $\mathbf{a}=\sum_{i=1}^{d}\alpha_i\mathbf{e}_i$ that yields the largest MI between f and p_a must be optimized iteratively. The optimization procedure for the MI was implemented as follows. The initial guess $\mathbf{a}_0$ for $\mathbf{a}$ is generated randomly, or corresponding to the optimized Pearson coefficient (eqs. (5)). Subsequently, $I(f,p_a)$ is optimized via a sequence of rotations $\mathbf{a}_{\ell+1}=\mathbf{R}_\ell\mathbf{a}_\ell$ which effect only three coefficients α_{i_ℓ}, α_{j_ℓ}, and α_{k_ℓ}. Hence,

$$\begin{aligned}&(\alpha_1,\ldots,\alpha_{i_{\ell+1}},\ldots,\alpha_{j_{\ell+1}},\ldots,\alpha_{k_{\ell+1}},\ldots,\alpha_d)\\&=\mathbf{R}_\ell(\alpha_1,\ldots,\alpha_{i_\ell},\ldots,\alpha_{j_\ell},\ldots,\alpha_{k_\ell},\ldots,\alpha_d)\end{aligned}\tag{13}$$

where $\alpha_{i_{\ell+1}}$, $\alpha_{j_{\ell+1}}$, $\alpha_{k_{\ell+1}}$ are derived from the α_{i_ℓ}, α_{j_ℓ}, α_{k_ℓ} by a three-dimensional (3D) rotation $\mathbf{R}_\ell^{3D}$,

$$(\alpha_{i_{\ell+1}},\alpha_{j_{\ell+1}},\alpha_{k_{\ell+1}})=\mathbf{R}_\ell^{3D}(\alpha_{i_\ell},\alpha_{j_\ell},\alpha_{k_\ell}).\tag{14}$$

For each optimization step, the i_ℓ, j_ℓ, and k_ℓ are randomly chosen from the d dimensions.

Each 3D rotation matrix $\mathbf{R}_\ell^{3D}$ is chosen such that it optimizes the MI in the $i_\ell-j_\ell-k_\ell$ subspace. To optimize $\mathbf{R}_\ell^{3D}$, approximately $n_p=150$ points are uniformly distributed on a unit sphere, each point corresponding to a possible 3D rotation with rotation angles ϕ_i and θ_i ($i=1,\ldots,n_p$). The MI $I_i(\phi_i,\theta_i)$ is computed for each of the n_p rotations. Subsequently, a set of spherical harmonics (up to order 5) is fitted to the n_p discrete $I_i(\phi_i,\theta_i)$, yielding a continuous and smoothed estimate of the MI as a function of the rotation angles ϕ and θ. The fit was implemented as a least-square fit using singular value decomposition. Eventually, the function $I(\phi,\theta)$ is optimized by Powell's method [24], yielding the best 3D rotation matrix $\mathbf{R}_\ell^{3D}$ of the ℓ^{th} optimization step. The 3D rotations are repeated in different $i_\ell-j_\ell-k_\ell$ subspaces until convergence.

The MI $I(f,p_a)$ is estimated from the discrete data sets by a binning procedure. Accordingly, the probability distributions $P_1(f')$ and $P_2(p'_a)$ are approximated by counting the occupancies $n_{f,i}$ and $n_{p,i}$ of $f(t)$ and $p_a(t)$, respectively, in bins $i=1,\ldots,N_b$. For the present study, the number of bins N_b was found to have a minor effect on the results. A reasonable choice was $N_b=50$. Likewise, $P(f',p'_a)$ are approximated by a two-dimensional (2D) binning, yielding the 2D occupancy $n_{fp,ij}$. The MI is estimated via

$$I(f,p_a)\approx\sum_{i,j=1}^{N_b}n_{fp,ij}/N_t\log\left(\frac{n_{fp,ij}}{n_{f,i}n_{p,i}/N_t}\right)\Delta f\,\Delta p_a,\tag{15}$$

where Δf and Δp_a denote the bin widths of $n_{f,i}$ and $n_{p,i}$, respectively. Note that the technique proposed here does not require an estimate of the absolute MI, but only of the relative change in MI due to a rotation $\mathbf{R}_\ell^{3D}$. Sophisticated and computationally demanding methods such as kernel density estimates are therefore unnecessary.

Simulation Setup

The Fs_{21} helix (originally introduced by Lockhart *et al.* [25], Sequence Ace-A_5[AAARA]$_3$A-NH_2) was modeled with PyMol [26]. The structures of T4 lysozyme (T4L) and leucine-binding protein (LBP) were taken from the protein data bank (PDB codes 256L [27] and 1USG [28], respectively). Likewise, the first structure in the NMR ensemble derived by Neidigh *et al.* (PDB code 1L2Y [29]) was used as initial structure for the simulations of Trp-cage. The Fs_{21}, T4L, Trp-cage, and LBP structures were placed into dodecahedral simulation boxes and solvated with 8828, 8479, 3042, and 17581 explicit water molecules, respectively. All simulation systems were neutralized by adding chloride ions. In addition, 150 mM sodium chloride was added to the Trp-cage and LBP systems.

The Fs_{21} helix and LBP were simulated with the AMBER03 [30] force field and the TIP3P water model was applied [31]. Ion parameters were taken from Smith *et al.* [32]. Trp-Cage and T4 lysozyme were simulated with the OPLS all-atom force field [33] and the TIP4P water model [34]. All simulations were carried out using the GROMACS simulation software [35,36]. Electrostatic interactions were calculated at every step with the particle-mesh Ewald method [37,38]. Short-range repulsive and attractive dispersion interactions were described by a Lennard-Jones potential, which was cut off at 1.0 nm (0.8 nm for the AMBER03 simulation). The SETTLE [39] algorithm was used to constrain bond lengths and angles of water molecules, and LINCS [40] was used to constrain the peptide bond lengths, allowing a time step of 2 fs.

The temperature in the Fs_{21} and T4L simulations was kept constant by weakly ($\tau=0.1$ ps) coupling the system to a temperature bath [41] of 300 K. Likewise, the pressure was kept constant by weakly coupling the system to a pressure bath of 1bar with a coupling constant τ of 1 ps. The LBP and Trp-cage systems were coupled to a Nosé-Hoover thermostat [42,43] ($\tau=2$ ps) at 300 K and 400 K (to trigger unfolding), respectively, and the pressure was kept at 1 bar using the Parrinello-Rahman pressure coupling scheme [44] ($\tau=5$ ps).

The Fs_{21} helix was simulated for 250 ns and the structure was written to the hard disk every picosecond. During the simulation the helix partially unfolded and refolded for a number of times. Because we chose to consider collective motions of an intact helix only, the secondary structure was determined for every frame with DSSP [45] and only frames with a complete helix were used for further analysis. Approximately 53.000 such frames were found which were combined into one 'trajectory' of 53 ns with time step 1 ps. The lysozyme simulation system was simulated for 460 ns, and the LBP system for 100 ns. The Trp-cage protein was simulated 8 times for 40 ns with different initial velocities. The volume of the catalytic cleft of T4L was estimated as explained and illustrated in supporting Figure S1.

Results

In the following we apply FMA on four biological examples, and demonstrate how quite different functional quantities f can be related to collective protein motions. In the first three examples of increasing complexity (Fs_{21} helix, T4 lysozyme, and Trp-cage) the Pearson coefficient turned out to be sufficient to detect correlations between the respective functional quantity and collective motions. With the final example (leucine-binding protein) we demonstrate how the MI can elucidate correlation in cases where the Pearson coefficient fails.

As a first and trivial example we analyze collective motions related to the end-to-end distance of the Fs_{21} helix. The example (including figures) is presented in supporting Text S2 and illustrates the application of FMA in some detail. Because the PCA vectors correspond to the harmonic modes of a simple helical spring, the decomposition of the end-to-end distance into contributions from different PCA modes is particularly instructive in this example.

Collective Motions of T4 Lysozyme Involved in Enzymatic Activity

Domain motions of hen lysozyme have been proposed more than 30 years ago [46,47]. Likewise, domain motions in T4 lysozyme (T4L) have been studied intensively by X-ray crystallography [27,48–50], site-directed spin labeling [51], as well as by theoretical approaches such as normal mode analysis, MD and PCA [47,52–54].

Here we demonstrate how FMA can be applied to determine the collective motions which are putatively involved in the enzymatic activity of T4L. Two functional quantities $f(t)$ related to the enzymatic activity are considered for the analysis. (i) The volume of the catalytic cleft V_{cleft}, highlighted as red surface in Fig. 2A, and (ii) the distance d_{ED} between the carboxyl groups of the catalytic residues Asp20 and Glu11 (Fig. 2B). The volume V_{cleft} is biologically significant because opening and closure of the cleft is expected to be involved in substrate binding and product release. The distance d_{ED} is a direct measure of the geometry of the catalytic site. According to the textbook mechanism proposed by Phillips [55], Glu11 protonates the glycosidic oxygen while Asp20 stabilizes the produced oxocarbenium ion intermediate. Hence, the carboxyl groups of Glu11 and Asp20 must simultaneously arrange closely to the glycosidic bond. d_{ED} is therefore an easily assessable observable that probes enzymatically active configurations. The distance between the carboxyl groups was measured as the distance between the C_δ atom of Glu11 and the C_γ atom of Asp20.

In the following, the results of the FMA of V_{cleft} and d_{ED} are presented in a relatively compact fashion. For a more detailed presentation of FMA we refer to the illustrative **α-helix** example in supporting Text S2. In a first step, the basis set $\{\mathbf{e}_i\}$ was derived by a PCA on the backbone atoms, using the 460-ns T4L simulation. The first 20 PCA vectors were used as basis set for the FMA. The motions along the first three PCA vectors are shown in Fig. 2C. The first PCA vector corresponds to the well-studied hinge-bending mode of T4L [46,51], and the second to a twisting mode, mainly characterized by a rotation of the smaller (N-terminal) lysozyme domain. The third PCA vector corresponds to the torsion of the N-terminal domain with respect to the C-terminal domain.

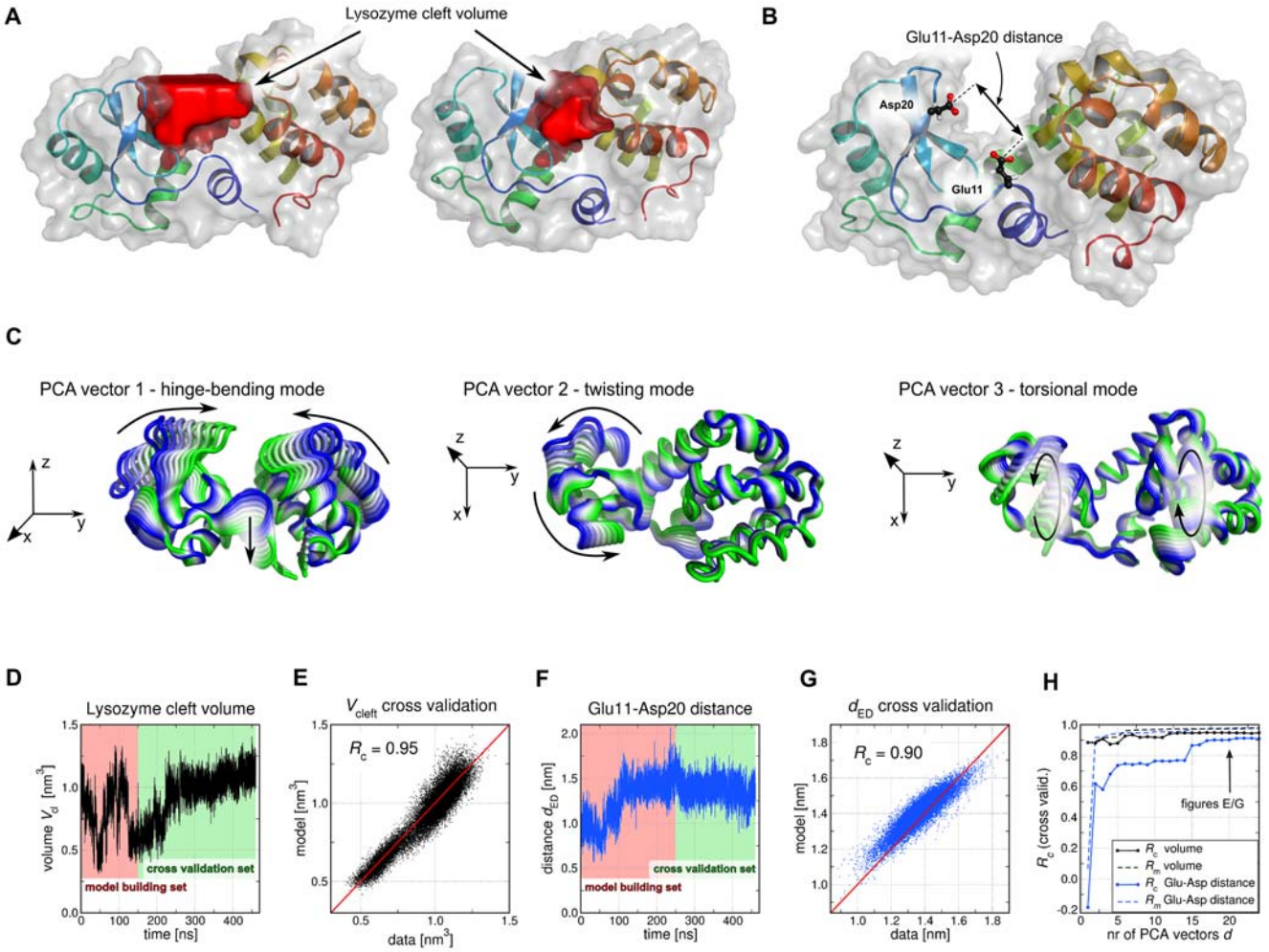

Figure 2. Functional mode analysis of catalytic cleft volume V_{cleft} and Glu11-Asp20 distance d_{ED} of T4 lysozyme (T4L). (A/B) T4L in cartoon and surface representation. The catalytic cleft is shown as red surface (A), and Glu11 and Asp20 are depicted in ball-and-stick representation (B). (C) The motions along the first three PCA vectors. (D) V_{cleft}, and (F), d_{ED} versus the simulation time (black and blue curves, respectively). The first 180 ns (250 ns for d_{ED}) were used as model building sets (red background), the remaining simulation frames as cross-validation sets (green background). (E/G) Scatter plots of the data versus the model using the cross-validation sets only. (H) Correlations R_m and R_c for V_{cleft} (black curves) and d_{ED} (blue curves), presented as a function of the number of principal components d used during the optimization.
doi:10.1371/journal.pcbi.1000480.g002

In Figures. 2D and 2F, V_{cleft} and d_{ED} are plotted as a function of simulation time, respectively. The first 180 ns (250 ns for d_{ED}) were applied as model building sets (red background), the remaining frames as cross-validation sets (green background). For both V_{cleft} and d_{ED}, the respective collective vector **a** was optimized by maximizing R, yielding linear models for V_{cleft} and d_{ED}. The models are validated in figs. 2E and 2G, showing scatter plots between simulation data and models using the respective cross-validation sets only. Strong cross-validated correlation ($R_c = 0.95$ and 0.90, respectively) between model and data was found, confirming the validity of the models. (The corresponding scatter plots using the model building sets are presented in Figure S3.) Note that side chain fluctuations of Glu11 and Asp20 cannot be modeled from backbone PCA modes. Yet the derived model for d_{ED} favorably correlates with the data ($R_c = 0.90$), indicating that side chain fluctuations have only a minor effect on d_{ED}. Figure 2H shows the R_c values for V_{cleft} and d_{ED} as black and blue curves, respectively, as a function of the number of PCA vectors d used in the FMA. Apparently, the first PC already provides a good model for V_{cleft} ($R_c = 0.89$). In contrast, at least 15 PCs are required to construct a good model ($R_c > 0.85$) for d_{ED}.

The convergence of FMA of V_{cleft} with the number of frames in the model building set is analyzed in Fig. 3. The figure plots R_c and R_m between the V_{cleft} data and V_{cleft} model as a function of the simulation time in the model building set. All remaining frames of the 460-ns trajectory were applied for cross validation. Remarkably, using only 10 ns for model building yields a reasonable model ($R_c > 0.85$) for the remaining frames. In contrast, using less than 0.5 ns for model building yields a highly overfitted model, as visible from the large R_m as compared to R_c. The related analysis for the FMA of d_{ED} is presented in Fig. S2B.

Figure 4 presents the collective vectors related to V_{cleft} and d_{ED}, as well as the contributions of different PCs to V_{cleft} and d_{ED}. The results for V_{cleft} are presented as black bars and curves, the results for d_{ED} in blue. The coefficients α_i of **a** (or β_i of the linear model, eq. (6)), are shown in Fig. 4A. Note that β_i quantifies the effect of the i^{th} PC on V_{cleft} (or d_{ED}) *per nanometer displacement* in PCA space. Because the variance of the PCs rapidly decay with increasing PC index i (Fig. 4B), only the first PCs substantially contribute to the variances of V_{cleft} and d_{ED} (Fig. 4C). Remarkably, the first PC almost completely accounts for $\text{var}(V_{\text{cleft}})$, whereas the second PC accounts for $\text{var}(d_{\text{ED}})$. Figure 4D presents the cumulative contributions of the PCs to $\text{var}(V_{\text{cleft}})$ and $\text{var}(d_{\text{ED}})$ as derived from the respective models as solid curves, and the variances of $\text{var}(V_{\text{cleft}})$ and $\text{var}(d_{\text{ED}})$ as dashed curves. The plot confirms that the models indeed account for a large fraction the variances of V_{cleft} and d_{ED}, respectively.

Which are the ewMCMs contributing to V_{cleft} and d_{ED}? Applying eq. (12) shows that the ewMCMs contributing to V_{cleft} and d_{ED} are highly related to PCA vectors 1 and 2, respectively, a finding in agreement to Fig. 4C. For the illustration of these motions we therefore refer to the PCA vectors depicted in Fig. 2C. In addition, the MCM and the ewMCM for both V_{cleft} and d_{ED} are shown in supporting videos S1 and S2.

Taken together, the FMA provides a comprehensive picture of the collective motions involved in the catalytic activity of T4L. The hinge-bending mode (Fig. 2C) dominates closing and opening of the catalytic cleft, presumably facilitating substrate binding and release. Surprisingly, this mode leaves the active site geometry virtually unaffected. In contrast, the twisting mode dominates the distance d_{ED} between the carboxyl groups of Glu11 and Asp20. Hence, a major collective rotation of the N-terminal domain with respect to the C-terminal domain may be required to position Glu11 and Asp20 into an enzymatically active configuration.

Initial Unfolding of Trp-Cage

Trp-cage is a 20-residue miniprotein designed by Neidigh *et al.* [29]. With a folding time of 4 μs [56] Trp-cage is the fastest folding

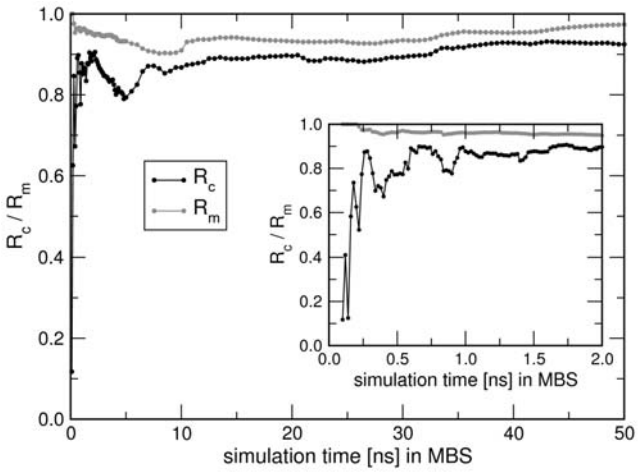

Figure 3. Convergence of FMA of the lysozyme cleft volume V_{cleft}. R_m (gray) and R_c (black curve) as a function of simulation time in the model building set (MBS). For each data point in the figure, all remaining frames from the 460-ns trajectory were used as cross validation set, and the first 20 PCA vectors were applied as the basis set to construct **a**. The inset displays the simulation time in the MBS is in a detailed scale. Applying approx. 10 ns of simulation as MBS is sufficient to yield a reasonable model ($R_c > 0.85$) for the remining frames. Applying less than 0.5 ns as MBS yields an highly overfitted model.
doi:10.1371/journal.pcbi.1000480.g003

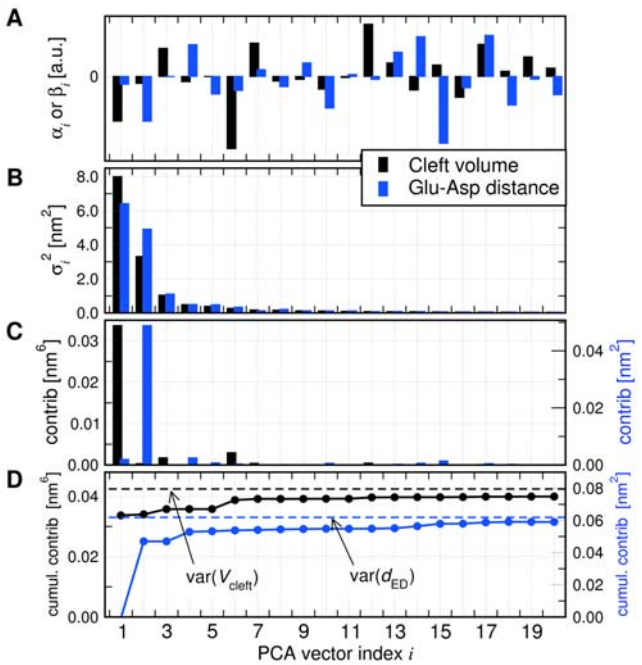

Figure 4. Collective vector 'a' related to the lysozyme cleft volume V_{cleft} (black) and to the distance d_{ED} between Glu11 and Asp20 (blue). (A) components α_i of **a** with respect to the PCA vectors $\mathbf{e}_i$. (B) Variances σ_i^2 of the principal components (PCs), (C) contribution of the i^{th} PC to the variance of the model, and (D) the cumulative contribution of principal component i to the variance of the model. The dashed lines indicates the variances of V_{cleft} and d_{ED} during the simulation.
doi:10.1371/journal.pcbi.1000480.g004

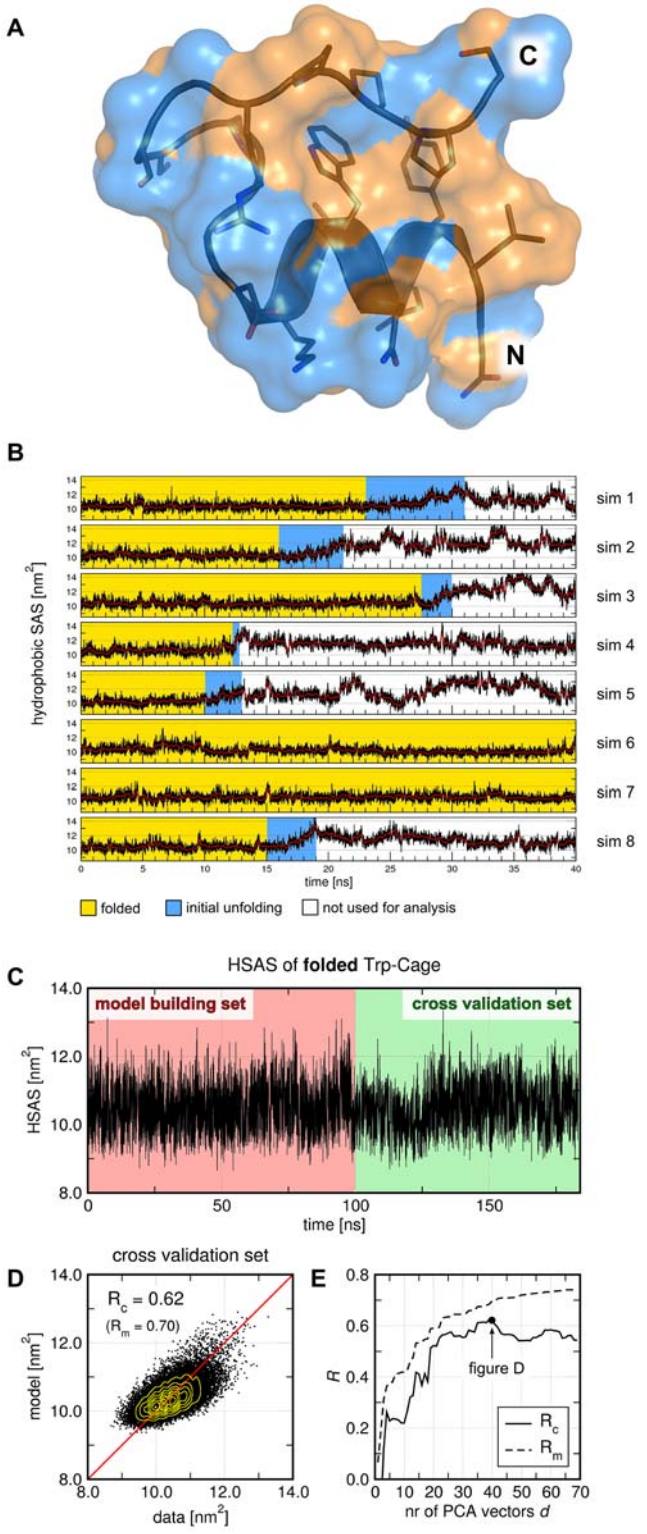

Figure 5. Functional mode analysis of the hydrophobic solvent-accessible surface (HSAS) of Trp-cage in the folded state. (A) Trp-cage protein in the folded state, shown in cartoon and surface representation. The HSAS is shown as orange surface, the hydrophilic SAS as blue surface. (B) HSAS during 8 independent simulations (sim 1-8). HSAS (black curves), and to guide the eye, the HSAS smoothed by a moving average (red curves). In simulation frames highlighted by a yellow background, Trp-cage was considered as folded. The initial unfolding events are highlighted by a blue background. (C) HSAS versus simulation time (black curve) combined from all folded states of the 8 Trp-cage simulations (B). The first 100 ns were used as model building set (red background), the remaining 83.3 ns as cross-validation set (green background). (D) Scatter plot of the data versus the model using the cross-validation set. (E) Correlations R_m and R_c of the model building and cross-validation sets, respectively, versus the number of PCA vectors d used during the optimization.
doi:10.1371/journal.pcbi.1000480.g005

protein currently known. Trp-cage is characterized by a central tryptophan side chain (Trp6) which is surrounded by an α-helix (residues 2–8), a 3_{10}-helix (residues 11–14), and a C-terminal polyproline helix (Fig. 5A). Here we use Trp-cage as a model system to demonstrate how FMA can be applied to study the structural determinants underlying the initial unfolding process of a protein. To this end, the hydrophobic solvent-accessible surface area (HSAS) is used to quantify the state of unfolding.

Compared to the distances and volumes considered so far as functional quantities $f(t)$, explaining the HSAS by single collective mode is challenging. The HSAS is subject to strong noise and is a non-linear function of the atom coordinates. Only the linear parts of the dependence of the HSAS on the PCs is expected to be successfully captured by the linear model of eq. (6). The non-linear dependence on the PCs (that would have to be described as cross terms of the PCs) will appear as a noisy deviation from the model.

We use FMA to determine the collective motions related to the change in the HSAS, and hence, to the initial unfolding of Trp-cage. Three questions are addressed: (i) To which extent can a model based on a single collective motion explain a highly non-linear quantity such as the HSAS? (ii) Which collective motions increase the HSAS and, hence, represent the initial unfolding of Trp-cage? And (iii), can a model derived only from fluctuations in the folded state predict the HSAS during an unfolding event? To observe multiple unfolding events, eight 40-ns simulations were performed at a temperature of 400 K. The HSAS during the eight simulations is shown in Fig. 5B. In six of the eight simulations, the Trp-cage unfolded after simulation times between 10 and 27.5 ns (blue background in Fig. 5B). In the other two simulations the Trp-cage remained folded for the complete simulation time. From the eight simulations, all frames of the folded Trp-cage (yellow background in Fig. 5B) were combined into one 'folded trajectory' of 183.3 ns (916503 frames).

The HSAS of the 'folded trajectory' is plotted in Fig. 5C. The first 100 ns were applied as model building set in the FMA (red background), the remaining 83.3 ns as cross-validation set (green background). The basis set for the FMA was taken from a PCA of all heavy atoms (after a least-square fit of the backbone atoms onto the 1L2Y structure). The first 40 PCA vectors were used as basis set. The PCA vectors are not visualized because they do not correspond to easily interpretable motions. From the model building set, a linear model for the HSAS was derived using eqs. (7), and the model was validated using the cross-validation set (Fig. 5D). The correlation between data and model is substantially weaker ($R_c = 0.62$, $R_m = 0.70$, see Fig. S3D) than in the previous examples. As expected, the HSAS *in the folded state* is only partially captured by the linear model. To reach a similar model quality as in the previous examples, additional non-linear cross terms of the PCs would have to be included into the model. Without such additional terms, the deviation from the linear model appears as noise. Analysis of the difference between data and model shows that the noise is normally distributed around zero with a standard deviation of 0.36 nm^2 (not shown).

To avoid overfitting, R_m and R_c are plotted as a function of the number of PCA vectors d used as basis set (Fig. 5E). Both R_m and R_c increase up to $d = 40$, corresponding to an improvement of the

model. For $d>40$, only R_m increases, but R_c decreases with d. Hence, using more than 40 PCA vectors as basis set would yield an overfitted model.

The ewMCM related to the increase in HSAS is visualized in Fig. 6A. The motion is mainly characterized by a lift-off motion of the polyproline helix with respect to the α-helix. The ewMCM and the MCM along **a** are also shown in video S3. The components α_i of **a** (or β_i of the model) are shown in Fig. 6B, and the variances of the PCs in Fig. 6C. As visible from Fig. 6D, many PCs contribute to the variance of the HSAS var(HSAS). Remarkably, PCs with larger index i substantially contribute the the HSAS, although they hardly contribute to the MSF of the atom positions (compare Fig. 6C). The cumulative contribution of the PCs to var(HSAS) (Fig. 6E) indicates that approximately 50% of the variance (corresponding to 70% of the standard deviation) of the HSAS in the folded state are explained by the linear model.

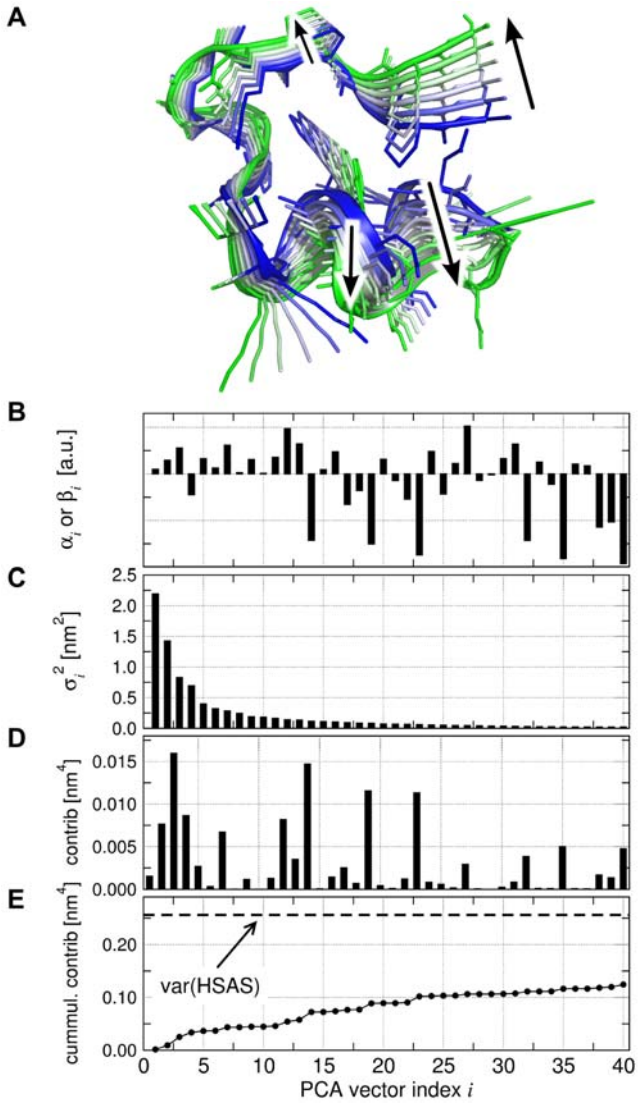

Figure 6. Collective motion related to the increase in hydrophobic solvent-accessible surface (HSAS) of Trp-cage. (A) Cartoon representation of the ensemble-weighted MCM contributing to the increase in HSAS. Side chains are shown as sticks. (B) Eigenvalues σ_i^2 of the principal components (PCs), (C) components α_i of the collective vector **a** with respect to the PCA vectors $\mathbf{e}_i$. (D) Contribution and (E) cumulative contribution of the i^{th} PC to the variance of the model. The dashed line indicates the variance var(HSAS) of the HSAS during the simulation.
doi:10.1371/journal.pcbi.1000480.g006

Can the model for the HSAS derived from fluctuations *in the folded state* predict the HSAS during unfolding? To assess this particularly rigorous test for the validity of the model, the HSAS during six independent unfolding events was monitored (blue background in Fig. 5B). Figure 7A displays two examples for the HSAS during unfolding events as black curves, and the HSAS predicted by the model as gray curves. The corresponding plots for all six unfolding events are shown in Fig. S4. Good agreement is found with correlation coefficients R between data and model in the range of 0.72 to 0.88 (insets in Fig. 7 and S4). The respective scatter plot of HSAS data versus model as combined from all six unfolding events is shown in Fig. 7B. Reasonable agreement ($R=0.81$) between data and model is found. Note that such unfolding events are not present the model building set. Hence, the model derived only from the folded state displays predictive power during a process (initial unfolding) which did not occur in the model building set. Noteworthy, favorable correlation between data and model during unfolding can only be achieved if the collective unfolding modes are at least partially sampled in the folded state fluctuation. If unfolding modes do not sufficiently fluctuate in the folded state, no correlation between such modes and the HSAS can be detected. Figures 7 and S4 show, however, that folded state fluctuations in this case are sufficient to construct the HSAS model (and hence, the functional mode) that holds during initial unfolding.

A Non-Linear Example: RMSD of Leucine-Binding Protein

As an example for a non-linear correlation between a functional quantity $f(t)$ and collective motions we consider the root mean square deviation (RMSD) of the backbone atoms of l-leucine-binding protein (LBP). LBP is a two-domain transport protein (Fig. 8A) that is subject to a large hinge motion (0.7 nm RMSD) upon ligand binding [28,57]. The RMSD was computed with respect to the (apo) crystal structure. The RMSD increases as the protein deviates from the reference, irrespective of the direction of the collective motion. Hence, it can not be explained in terms of a linear function of a collective coordinate. Because the RMSD is frequently assessed in MD studies, we use it as a model quantity to demonstrate the use of mutual information (MI) in FMA.

The RMSD is plotted in Fig. 8B. The collective vector **a** was optimized such that $p_a(t)$ displays the maximal MI to the RMSD. To this end, the first 80 ns of the simulation were applied as model building set, omitting the first nanosecond for equilibration (red background in Fig. 8B), and the remaining frames were applied as cross-validation set (green background). The first 10 PCA vectors from a PCA of the backbone atoms were used as basis set for **a** (not shown). Figure 8C presents the RMSD versus the optimized collective coordinate p_a for the model building set (red dots) and the cross-validation set (green dots). As expected, the RMSD and p_a are substantially correlated, and the correlation is highly non-linear as visible from the non-linear RMSD-p_a point cloud. The non-linear model for the RMSD was constructed by fitting a cubic B-spline (black curve in Fig. 8C) to the RMSD-p_a points of the model building set. Using this model, the RMSD of the cross-validation set was predicted and compared to the RMSD from the simulation (red dots in Fig. 8D). Excellent agreement ($R_c=0.97$) between data and model is found.

For comparison, we tried to derive a linear model for RMSD using the Pearson coefficient R as correlation measure. However, this linear model has little predictive power as visible from model-versus-data scatter plot of the cross-validation set (black dots in Fig. 8D, $R_c=0.42$). The failure of R to detect the correlation

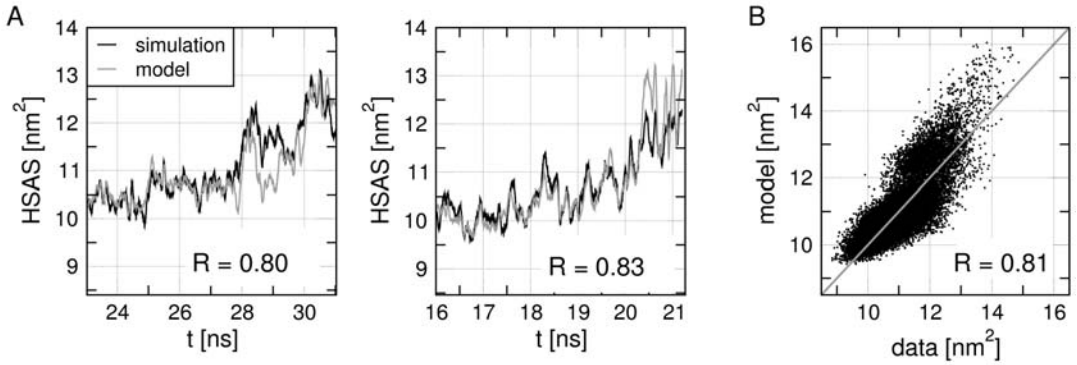

Figure 7. Predictive power of HSAS model during the initial unfolding events. (A) HSAS (black) during unfolding events of Trp-cage simulations 1 and 2 and the prediction by the HSAS model (gray). The correlations R between model and data are printed as insets. To facilitate the comparison between data and model in these plots, all HSAS curves were slightly smoothed by running averages. The R-values were computed from the non-smoothed data (not shown). The HSAS and the prediction of all six unfolding events are shown in the supporting material. (B) Scatter plot of HSAS data versus model as collected from all six unfolding events.
doi:10.1371/journal.pcbi.1000480.g007

between RMSD and a single collective motion is also apparent from Fig. 8E which plots R_m and R_c as a function of the number of PCA vectors d used as basis set. Irrespective of d, R_c (black curve) is substantially smaller than R_m (blue curve), indicating overfitting. In this example, using more than 10 PCA vectors as basis set would increase R_c, but R_c remains much smaller than R_m. Note that the model derived using the MI as correlation measure displays excellent correlation between data and model for both model building and cross-validation sets (green and red curves in Fig. 8E, respectively).

Figure 9F presents the analysis of the convergence of FMA with the number of frames in the model building set by plotting R_c and R_m as a function of simulation time in the model building set. All remaining frames of the 100-ns trajectory were applied for cross validation, and the first 10 PCA vectors were used as basis set. When optimizing the MI (red and green curves), a model building set of 11 ns is sufficient to derive a good model ($R_c \geq 0.9$) for the remaining simulation. In contrast, using less than 5 ns for model building may yield a highly overfitted model, as visible from a

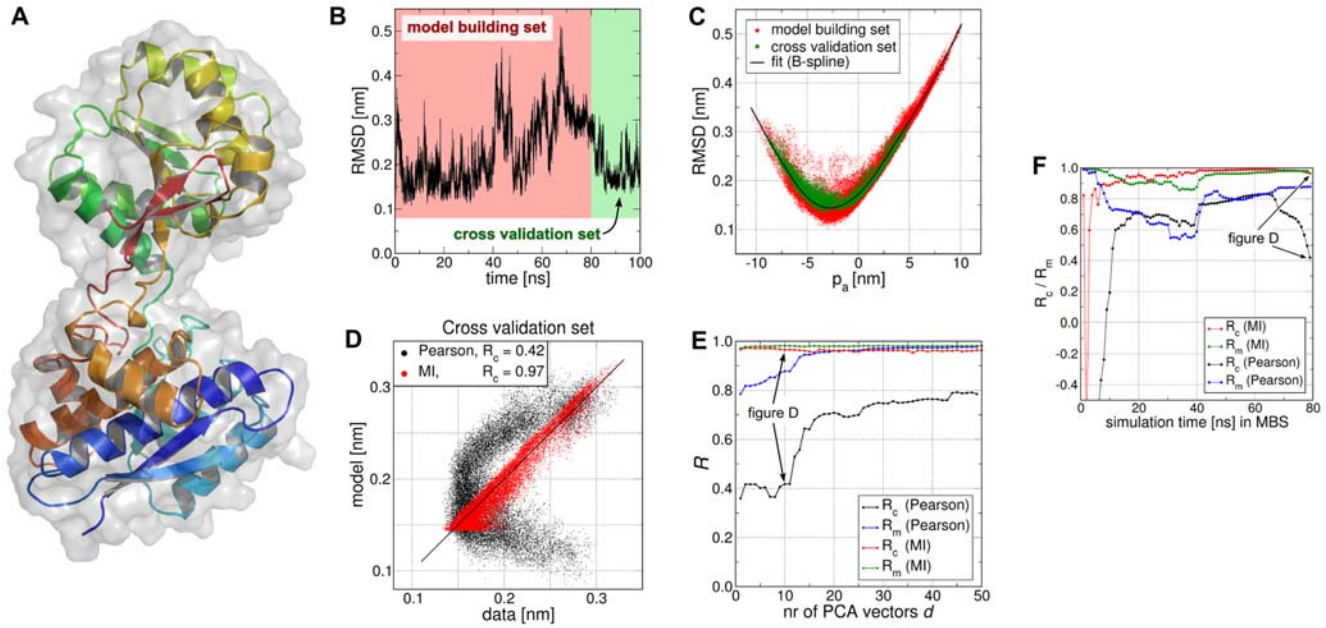

Figure 8. Functional mode analysis of the RMSD of leucine-binding protein (LBP) with respect to its apo structure. (A) Apo structure of LBP in cartoon and surface representation. (B) RMSD with respect to the apo structure versus simulation time. Model building and cross-validation sets are highlighted by red and green background, respectively. (C) RMSD versus the collective coordinate p_a optimized via mutual information (MI). Model building set (red dots), cross-validation set (green dots), and spline fitted to the model building set (black curve). (D) Scatter plot of cross-validation set showing model versus data. Optimizing MI yields favorable correlation (red dots, $R_c = 0.97$), optimizing the Pearson coefficient only poor correlation (black dots, $R_c = 0.42$). (E) Correlations R_m and R_c of the model building and cross-validation sets, respectively, versus the number of PCA vectors d used during the optimization. MI optimization (green/red curves) is compared to Pearson optimization (blue/black curves). (F) Convergence of FMA with simulation time in the model building set (MBS). R_m and R_c (from Pearson and MI optimization, compare legend) are shows as a function of simulation time in the MBS using 10 PCA vectors as basis set. All remaining frames of the 100-ns trajectory were applied as cross validation set.
doi:10.1371/journal.pcbi.1000480.g008

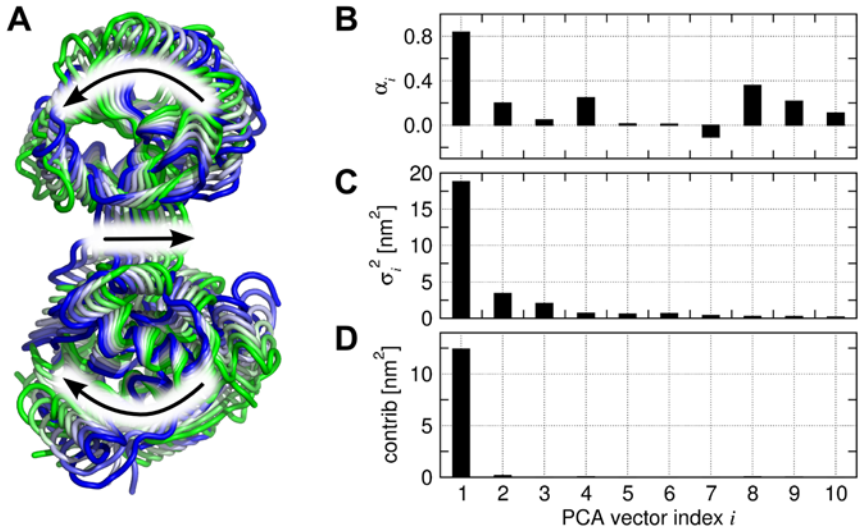

Figure 9. Collective motion related to the increase in RMSD of leucine-binding protein (LBP). (A) Backbone representation of the ensemble-weighted MCM motion contributing to the increase in RMSD. (B) components α_i of the collective vector **a** with respect to the PCA vectors $\mathbf{e}_i$. (C) Eigenvalues σ_i^2 of the principal components (PCs), (D) Contribution of the i^{th} PC to the variance of the collective coordinate p_a.
doi:10.1371/journal.pcbi.1000480.g009

small (or even negative) R_c (red curve). When optimizing the Pearson coefficient (black and blue curves), the model quality as measured from R_c is substantially poorer as compared to the model optimized via MI. In addition, R_c is highly depdendent on the length of the model building set, further emphasizing that the the Pearson coefficient is an unsuitable measure to assess correlation between the RMSD and a single collective mode.

The ewMCM that effects the optimized collective coordinate p_a (and hence, the RMSD) is visualized in Fig. 9A. The motion is characterized by a large hinge of the two domains with respect to each other. The collective motion is decomposed into the PCs in figs. 9B–D. The coordinates α_i of **a** are sown in Fig. 9B and the variances of the PCs in Fig. 9C. Figure 9D displays the contributions of the PCs to the variance of p_a, indicating that the first PC almost completely accounts for the variance of p_a. This finding is expected because the first PC is constructed such that it accounts for the largest possible fraction of the RMSD. Hence, Fig. 9D may be considered as a further validation of the technique.

Discussion

We have presented a novel analysis technique termed functional mode analysis (FMA) to systematically identify functionally relevant collective motions in proteins dynamics. Given an arbitrary quantity $f(t)$ considered relevant to the function of the protein of interest, the approach extracts the linear collective motion which is maximally correlated to $f(t)$. We have used two different measures to quantify the correlation between $f(t)$ and the collective motion. (i), the Pearson coefficient which measures linear correlation only, and (ii), the mutual information (MI) which can assess any kind of correlation including non-linear and higher order correlation. The 'maximally correlated motion' (MCM) must be distinguished from the 'ensemble-weighted MCM' (ewMCM) that –based on the sampling in the input ensemble– has the largest likelihood to contribute to large fluctuations in $f(t)$. Numerous proteins accomplish their biological function by structural transitions such as hinge motions, domain rotations, or side chain reorientations [2]. For many proteins it is far from obvious how functionally relevant quantities are related to such collective atomic motions. In such cases, the proposed technique is expected to elucidate relations between collective motions (i.e. protein dynamics) and protein function. Moreover, the MCM or the ewMCM can be enhanced or steered during a follow-up simulation, allowing one to trigger functional transitions or to enhance the sampling of rare functional events [58,59]. Alternatively, the MCM can be biased to compute free energy differences between different functional states (using, e.g., umbrella sampling).

The success of the technique rests on two prerequisites: (i) The collective motion **a** must be representable by a linear combination of the chosen basis set $\{\mathbf{e}_i\}$. At the same time, the basis set should not be too large to avoid overfitting. (ii) Sufficient (linear or non-linear) correlation between $f(t)$ and the single collective motion must be detectable. Future efforts will focus on ameliorating the first condition.

Optimization of the collective vector **a** via the Pearson coefficient is closely related to the construction of a model for $f(t)$ as a linear function of a set of given collective coordinates. When using MI as correlation measure, a non-linear model for $f(t)$ can be constructed from a general set of functions (such as splines). The given collective coordinates used as basis set may correspond to motions along PCA vectors, normal modes, or to motions in any other useful coordinate system such as rotations of dihedral angles. If these given coordinates have an intuitive meaning (such as the hinge-bending mode in T4L), the derived model can quantify the contributions of intuitive collective motions to the variance of $f(t)$, and hence, provide a functional interpretation of the collective dynamics.

The source code of an FMA implementation is available on the authors' web site http://www.mpibpc.mpg.de/groups/de_groot/software.html.

Supporting Information

Text S1 Text S1
Found at: doi:10.1371/journal.pcbi.1000480.s001 (0.06 MB PDF)

Text S2 FMA of the end-to-end distance of an α-helix
Found at: doi:10.1371/journal.pcbi.1000480.s002 (1.92 MB PDF)

Figure S1 Estimation of the volume of the lysozyme catalytic cleft. A block of test atoms with the approximate shape of the

binding site was set up by placing the test atoms on a grid of spacing 1 Å (red block in fig. S1A). The block was placed into the catalytic cleft of a reference structure of T4 lysozyme (T4L). Each T4L structure from the simulation trajectory was fitted onto the reference structure using a least square fit on the backbone atoms. Figure S1A shows one example of a fitted structure, together with the block of test atoms. Subsequently, the test atoms which overlapped with the fitted structure were removed with the genbox tool (fig. S1B). Every remaining atom contributed 1 $Å^3$ to the cleft volume.
Found at: doi:10.1371/journal.pcbi.1000480.s003 (0.27 MB JPG)

Figure S2 Convergence of FMA with simulation time. (A–C) Correlations R_c (black curves) and R_m (red curves) of the cross validation and model building sets, respectively, as a function of the number of frames (or simulation time) in the model building set (MBS). (A) Helix end-to-end distance L_h. Approximately 30 frames are sufficient to construct a reliable model for L_h, as visible from the R_c curve (compare Text S2). (B) T4 lysozyme Glu11-Asp20 distance d_{ED}. Approximately 10 ns of simulation are sufficient to yield reasonable models for V_{cleft} and d_{ED}, although the model quality as measured by R_c may slightly increase when applying more than 40 ns as MBS. (C) Hydrophobic solvent-accessible surface (HSAS) of Trp-cage in the folded state. From the folded states of the 8 Trp-cage simulations (yellow background in Fig. 5B), an increasing fraction (e.g. 20%) was used as model building set, whereas the remaining fraction (e.g. 80%) of the folded states were applied as cross validation set. As visible from the black R_c curve, applying more than 30% of the simulation hardly improves the prediction for the remaining frames. The respective plots for the lysozyme cleft volume and the RMSD of leucine-binding protein are shown in Figs. 3 and 8F, respectively.
Found at: doi:10.1371/journal.pcbi.1000480.s004 (0.06 MB PNG)

Figure S3 Scatter plots showing model versus data of the model building sets. (A) Helix end-to-end distance, (B) Glu11-Asp20 distance of T4 lysozyme (T4L), (C) cleft volume of T4L, (D) hydrophobic solvent-accessible surface of Trp-cage, and (E) RMSD of backbone atoms of leucine-binding protein (LBP) with respect to its apo structure. Optimization of the mutual information (MI, red dots) yields larger correlation than optimization of Pearson's coefficient (black dots).
Found at: doi:10.1371/journal.pcbi.1000480.s005 (0.26 MB PNG)

Figure S4 Predictive power of the model for the hydrophobic solvent-accessible surface (HSAS) during six initial unfolding events. HSAS (black curves) during unfolding events and the prediction for the HSAS by the model (red curves). Note that the model was derived only from fluctuations in the folded state. The correlation R between model and data (printed as insets) lies in the range of 0.72 to 0.88. To facilitate the comparison between data and model in these plots, all HSAS curves were slightly smoothed by running averages. The R-values were computed from the non-smoothed data (not shown).
Found at: doi:10.1371/journal.pcbi.1000480.s006 (0.12 MB PNG)

Video S1 Movie showing the MCM (top) and the ewMCM (bottom) related to the increase of cleft volume V_{cleft} of T4 lysozyme.
Found at: doi:10.1371/journal.pcbi.1000480.s007 (2.99 MB MPG)

Video S2 Movie showing the MCM (top) and the ewMCM (bottom) related to the increase of the distance d_{ED} between Glu11 and Asp20 in T4 lysozyme.
Found at: doi:10.1371/journal.pcbi.1000480.s008 (3.18 MB MPG)

Video S3 Movie showing the MCM (left) and the ewMCM (right) related to the increase in hydrophobic solvent-accessible surface of Trp-cage.
Found at: doi:10.1371/journal.pcbi.1000480.s009 (2.35 MB MPG)

Acknowledgments

We thank D. Matthes and M. Kubitzki for providing us with simulation trajectories of the Fs_{21} helix and of T4L, respectively. We are grateful to H. Grubmüller for valuable discussions, and to D. Matthes and H. Grubmüller for carefully reading the manuscript.

Author Contributions

Conceived and designed the experiments: JSH BLdG. Performed the experiments: JSH. Analyzed the data: JSH. Wrote the paper: JSH BLdG.

References

1. Henzler-Wildman K, Kern D (2007) Dynamic personalities of proteins. Nature 450: 964–972.
2. Gerstein M, Krebs W (1998) A database of macromolecular motions. Nucleic Acids Res 26: 4280–4290.
3. Pelupessy P, Ravindranathan S, Bodenhausen G (2003) Correlated motions of successive amide N-H bonds in proteins. J Biomol NMR 25: 265–280.
4. Mittermaier A, Kay LE (2006) New tools provide new insights in NMR studies of protein dynamics. Science 312: 224–228.
5. Bourgeois D, Royant A (2005) Advances in kinetic protein crystallography. Curr Opin Struct Biol 15: 538–547.
6. Michalet X, Weiss S, Jäger M (2006) Single-molecule fluorescence studies of protein folding and conformational dynamics. Chem Rev 106: 1785–1813.
7. Yang H, Luo G, Karnchanaphanurach P, Louie TM, Rech I, et al. (2003) Protein conformational dynamics probed by single-molecule electron transfer. Science 302: 262–266.
8. Berendsen HJ, Hayward S (2000) Collective protein dynamics in relation to function. Curr Opin Struct Biol 10: 165–169.
9. Kitao A, Hirata F, Gō N (1991) 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. J Chem Phys 158: 447–472.
10. Amadei A, Linssen ABM, Berendsen HJC (1993) Essential dynamics of proteins. Proteins: Struct Funct Genet 17: 412–425.
11. Kitao A, Gō N (1999) Investigating protein dynamics in collective coordinate space. Curr Opin Struct Biol 9: 143–281.
12. Garcia AE (1992) Large-amplitude nonlinear motions in proteins. Phys Rev Lett 68: 2696–2699.
13. Gō N, Noguti T, Nishikawa T (1983) Dynamics of a small globular protein in terms of low-frequency vibrational modes. Proc Natl Acad Sci USA 80: 3696–3700.
14. Brooks BR, Karplus M (1983) Harmonic dynamics of proteins: Normal modes and fluctuations in bovine pancreatic trypsin inhibitor. Proc Natl Acad Sci USA 80: 6571–6575.
15. Levitt M, Sander C, Stern PS (1983) The normal modes of a protein: Native bovine pancreatic trypsin inhibitor. Int J Quant Chem: Quantum Biology Symposium 10: 181–199.
16. Hayward S (2000) Normal mode analysis of biological molecules. In: Becker OM, MacKerell AD, Roux B, Watanabe M, eds. Computational Biochemistry and Biophysics. New York: Marcel-Dekker. pp 153168.
17. Bahar I, Erman B, Haliloglu T, Jernigan RL (1997) Efficient characterization of collective motions and interresidue correlations in proteins by low-resolution simulations. Biochemistry 36: 13512–13523.
18. Cover T, Thomas J (1991) Elements of information theory. New York: Wiley.
19. Lange OF, Grubmüller H (2008) Full correlation analysis of conformational protein dynamics. Proteins 70: 1294–1312.
20. Mardia KV, Kent JT, Bibby JM (1979) Multivariate Analysis. San Diego: Academic Press.
21. de Boor C (1994) A Practical Guide to Splines Springer.
22. Van Aalten DMF, Amadei A, Vriend G, Linssen ABM, Venema G, et al. (1995) The essential dynamics of thermolysin — confirmation of hinge-bending motion and comparison of simulations in vacuum and water. Proteins: Struct Funct Genet 22: 45–54.
23. Bishop CM (2008) Pattern Recognition and Machine Learning. Berlin: Springer.
24. Brent RP (1973) Algorithms for Minimisation Without Derivatives Prentice Hall.

25. Lockhart DJ, Kim PS (1992) Internal stark effect measurement of the electric field at the amino terminus of an alpha helix. Science 257: 947–951.
26. DeLano WL (2002) The PyMOL molecular graphics system. http://www.pymol.org.
27. Faber HR, Matthews BW (1990) A mutant T4 lysozyme displays five different crystal conformations. Nature 348: 263–266.
28. Magnusson U, Salopek-Sondi B, Luck LA, Mowbray SL (2004) X-ray structures of the leucinebinding protein illustrate conformational changes and the basis of ligand specificity. J Biol Chem 279: 8747.
29. Neidigh JW, Fesinmeyer RM, Andersen NH (2002) Designing a 20-residue protein. Nat Struct Biol 9: 425–430.
30. Duan Y, Wu C, Chowdhury S, Lee MC, Xiong G, et al. (2003) A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations. J Comput Chem 24: 1999–2012.
31. Mahoney ME, Jorgensen WL (2000) A five-site model for liquid water and the reproduction of the density anomaly by rigid, nonpolarizable potential functions. J Chem Phys 112: 8910–8922.
32. Smith D, Dang L (1994) Computer simulations of NaCl association in polarizable water. The Journal of Chemical Physics 100: 3757.
33. Kaminski GA, Friesner RA, Tirado-Rives J, Jorgensen WL (2001) Evaluation and reparametrization of the OPLS-AA force field for proteins via comparison with accurate quantum chemical calculations on peptides. J Phys Chem B 105: 6474–6487.
34. Jorgensen WL, Chandrasekhar J, Madura JD, Impey RW, Klein ML (1983) Comparison of simple potential functions for simulating liquid water. J Chem Phys 79: 926–935.
35. Lindahl E, Hess B, Van der Spoel D (2001) GROMACS 3.0: a package for molecular simulation and trajectory analysis. J Mol Model 7: 306–317.
36. Van der Spoel D, Lindahl E, Hess B, Groenhof G, Mark AE, et al. (2005) GROMACS: Fast, flexible and free. J Comp Chem 26: 701–1719.
37. Darden T, York D, Pedersen L (1993) Particle mesh Ewald: an N·log(N) method for Ewald sums in large systems. J Chem Phys 98: 10089–10092.
38. Essmann U, Perera L, Berkowitz ML, Darden T, Lee H, et al. (1995) A smooth particle mesh ewald potential. J Chem Phys 103: 8577–8592.
39. Miyamoto S, Kollman PA (1992) SETTLE: An analytical version of the SHAKE and RATTLE algorithms for rigid water models. J Comp Chem 13: 952–962.
40. Hess B, Bekker H, Berendsen HJC, Fraaije JGEM (1997) LINCS: A linear constraint solver for molecular simulations. J Comp Chem 18: 1463–1472.
41. Berendsen HJC, Postma JPM, DiNola A, Haak JR (1984) Molecular dynamics with coupling to an external bath. J Chem Phys 81: 3684–3690.
42. Nosé S (1984) A molecular dynamics method for simulations in the canonical ensemble. Mol Phys 52: 255–268.
43. Hoover WG (1985) Canonical dynamics: Equilibrium phase-space distributions. Phys Rev A 31: 1695–1697.
44. Parrinello M, Rahman A (1981) Polymorphic transitions in single crystals: A new molecular dynamics method. J Appl Phys 52: 7182–7190.
45. Kabsch W, Sander C (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22: 2577–2637.
46. McCammon JA, Gelin B, Karplus M, Wolynes PG (1976) The hinge bending mode in lysozyme. Nature 262: 325–326.
47. Brooks B, Karplus M (1985) Normal modes for specific motions of macromolecules: application to the hinge-bending mode of lysozyme. Proc Natl Acad Sci USA 82: 4995–4999.
48. Matthews BW, Remington SJ (1974) The three dimensional structure of the lysozyme from bacteriophage T4. Proc Natl Acad Sci USA 71: 4178–4182.
49. Dixon MM, Nicholson H, Shewchuk L, Baase WA, Matthews BW (1992) Structure of a hingebending bacteriophage T4 lysozyme mutant, Ile3→Pro. J Mol Biol 227: 917–933.
50. Zhang XJ, Wozniak JA, Matthews BW (1995) Protein flexibility and adaptability seen in 25 crystal forms of T4 lysozyme. J Mol Biol 250: 527–552.
51. Mchaourab HS, Oh KJ, Fang CJ, Hubbell WL (1997) Conformation of T4 lysozyme in solution. hinge-bending motion and the substrate-induced conformational transition studied by site-directed spin labeling. Biochemistry 36: 307–316.
52. Hayward S, Kitao A, Berendsen HJC (1997) Model free methods of analyzing domain motions in proteins from simulation: A comparison of normal mode analysis and molecular dynamics simulation of lysozyme. Proteins: Struct Funct Genet 27: 425–437.
53. Hayward S, Berendsen HJC (1998) Systematic analysis of domain motions in proteins from conformational change: New results on citrate synthase and T4 lysozyme. Proteins: Struct Funct Genet 30: 144–154.
54. de Groot BL, Hayward S, van Aalten DMF, Amadei A, Berendsen HJC (1998) Domain motions in bacteriophage T4 lysozyme; a comparison between molecular dynamics and crystallographic data. Proteins: Struct Funct Genet 31: 116–127.
55. Phillips DC (1967) The hen egg white lysozyme molecule. Proc Natl Acad Sci U S A 57: 484–495.
56. Qiu L, Pabit SA, Roitberg AE, Hagen SJ (2002) Smaller and faster: the 20-residue Trp-cage protein folds in 4 micros. J Am Chem Soc 124: 12952–12953.
57. Penrose WR, Nichoalds GE, Piperno JR, Oxender DL (1968) Purification and properties of a leucine-binding protein from Escherichia coli. J Biol Chem 243: 5921–5928.
58. Amadei A, Linssen ABM, de Groot BL, van Aalten DMF, Berendsen HJC (1996) An efficient method for sampling the essential subspace of proteins. J Biom Str Dyn 13: 615–626.
59. Grubmüller H (1995) Predicting slow structural transitions in macromolecular systems: Conformational flooding. Phys Rev E 52: 2893–2906.