Partial least squares functional mode analysis: application to the membrane proteins AQP1, Aqy1 and CLC-ec1

Tatyana Krivobokova[1]
Institute for Mathematical Stochastics and Courant Research Center PEG,
Georg-August-University Göttingen, Germany

Rodolfo Briones[1]
Computational Biomolecular Dynamics Group,
Max-Planck Institute for Biophysical Chemistry, Göttingen, Germany

Jochen S. Hub
Computational Molecular Biophysics Group,
Dept. of Molecular Structural Biology,
Georg-August-University Göttingen, Germany

Axel Munk
Institute for Mathematical Stochastics,
Georg-August-University Göttingen and Statistical inverse problems group,
Max-Planck Institute for Biophysical Chemistry, Göttingen, Germany

Bert L. de Groot[2]
Computational Biomolecular Dynamics Group,
Max-Planck Institute for Biophysical Chemistry, Göttingen, Germany

[1]equal contribution

[2]Corresponding author. Address: Computational Biomolecular Dynamics Group, Max-Planck Institute for Biophysical Chemistry, Am Fassberg 11, 37077, Göttingen, Germany. Tel.: +(49)551-2012308, Fax: +(49)551-2012302

Abstract

We introduce a novel approach based on the recently introduced functional mode analysis to identify collective modes of internal dynamics that maximally correlate to an external order parameter of functional interest. Input structural data can be either experimentally determined structure ensembles or simulated ensembles, such as molecular dynamics trajectories. Partial least squares regression is shown to yield a robust solution to the multidimensional optimization problem, with a minimal and controllable risk of overfitting, as shown by extensive cross-validation. Several examples illustrate that the partial least squares based functional mode analysis successfully reveals the collective dynamics underlying the fluctuations in selected functional order parameters. Applications to T4 lysozyme, the Trp-cage, the aquaporin channels Aqy1 and hAQP1 and the CLC-ec1 chloride antiporter are presented in which the active site geometry, the hydrophobic solvent accessible surface, channel gating dynamics, water permeability (p_f), and a dihedral angle are defined as functional order parameters. The Aqy1 case reveals a gating mechanism that connects the inner channel gating residues with the protein surface, thereby providing an explanation of how the membrane may affect the channel. hAQP1 shows how the p_f correlates with structural changes around the aromatic/arginine region of the pore. The CLC-ec1 application shows how local motions of the gating Glu-148 couple to a collective motion that affects ion affinity in the pore.

Key words: principal component analysis; essential dynamics; molecular dynamics; Yeast-Aquaporin; human aquaporin-1; CLC chloride channel family

Introduction

Protein function frequently requires dynamics. Ranging from transporters to enzymes, from motors to signaling proteins, conformational transitions are usually at the heart of protein function. Consequently, a key step in understanding protein function is detailed knowledge of the underlying dynamics. Molecular dynamics (MD) simulations and related techniques are routinely used to study the dynamics of biomolecular systems at atomic detail at timescales of typically nanoseconds to microseconds. Although in principle allowing to directly address function-dynamics relationships, such analyzes are frequently hampered by the large dimensionality of a protein's configuration space, rendering it non-trivial to identify collective modes of motion that are directly related to a functional property of interest.

Principal component analysis (PCA) is a powerful tool to effectively reduce the dimensionality of a protein's configuration space (1, 2). Diagonalisation of the variance-covariance matrix yields a large number of eigenvectors with near-zero eigenvalues, corresponding to modes with only a minor contribution to the overall dynamics, leaving a relatively small percentage of principal modes that contribute to the vast majority of overall fluctuation. Even though PCA frequently aids a structural or functional interpretation enormously, it is not primarily designed for that purpose. PCA sorts the collective modes (eigenvectors) according to their contribution (eigenvalue) to the total mean-square fluctuation. Hence, the eigenvectors corresponding to the largest eigenvalues are defined as principal modes by virtue of the size of their fluctuation, irrespective of the actual contribution to a functional property of interest. Some functional properties may be influenced by the principal modes, but only in a specific combination, thereby further obscuring the relation between dynamics and function.

The PCA-based functional mode analysis (FMA) aims to overcome this problem by taking a linear combination of principal modes that is maximally correlated to a defined functional property of interest (3). FMA yields a linear model for an unidimensional functional property $f(t)$ (subsequently denoted as vector $\boldsymbol{f}$ to indicate that it is applicable to any ensemble, not only to time series). The correlation between the model and $\boldsymbol{f}$, particularly the correlation for a cross-validation subset of the data that was not used to train the model, provides a measure for the goodness of fit or the predictive power of the linear relation between coordinates (dynamics) and function. The linear model, also termed maximally correlated mode (MCM), can be expressed in terms of the original (cartesian) coordinates and visualized as a collective mode of motion, either directly or in an ensemble-weighted fashion (ewMCM for ensemble-weighted MCM), to yield a direct visualization of the dynamics underlying fluctuations in $\boldsymbol{f}$.

In its original implementation FMA uses the principal modes provided by PCA as a basis for the correlation optimization and therefore takes advantage of the dimensionality reduction provided by PCA, rendering it unnecessary to carry out the correlation optimization in the full coordinate space, that would easily lead to an overfitting issue due to the large number of parameters involved. However, this rests on the assumption that fluctuations in $\boldsymbol{f}$ are predominantly influenced by the principal modes, which may or may not be the case. Indeed, a number of FMA applications require a relatively high-dimensional PCA basis (3), indicating that not only principal modes contribute. Working with such a high-dimensional basis suffers from an inherent overfitting risk, and suggests that PCA does not offer an optimal basis in

such cases.

Here, we present a generalization of an FMA-based partial least squares (PLS) algorithm, that overcomes this issue by simultaneously optimizing model *and* basis. PLS-based FMA therefore yields a model with the lowest possible basis dimensionality that provides optimal correlation between fluctuations in $\boldsymbol{f}$ and protein dynamics. In other words the objective of the PLS algorithm is to get a relation (in this implementation a linear combination of coordinates) that correlate the best with $\boldsymbol{f}$, but at the same time that allows to identify the dynamics information in the input coordinates that contribute the most to the fluctuation in $\boldsymbol{f}$. Applications to the active-site geometry of T4 lysozyme and solvent accessible surface of the Trp-cage illustrate that a very specific combination of atomic fluctuations of the backbone atoms contribute to $\boldsymbol{f}$. Furthermore, FMA based on PLS requires a significantly smaller basis than PCA-based FMA. Due to the minimal dimensionality employed, the overfitting risk is minimized, leading to an optimal predictive power, as observed in cross-validation experiments. Furthermore, complex applications to the gating of aquaporin water channels, and to a CLC antiporter illustrate the use of PLS as a general and robust method to study function-dynamics relationships in proteins.

Theory

The partial least squares algorithm

A multiple linear least squares regression of the type $\boldsymbol{f} = \boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{\epsilon}$, with $\boldsymbol{f}$ a vector containing n samples of the unidimensional functional property to be described in terms of a $n\times p$ matrix $\boldsymbol{X}$ of p cartesian coordinates, yields a set of p-dimensional coefficients $\boldsymbol{\beta}$ with residuals $\boldsymbol{\epsilon}$. Minimizing $\boldsymbol{\epsilon}$ optimizes the *correlation* between $\boldsymbol{f}$ and the linear model $\boldsymbol{X}\boldsymbol{\beta}$. In practice, such a regression can yield poor results (particularly in prediction) if some of the columns of $\boldsymbol{X}$ are nearly dependent (or colinear in statistical terms).

In PCA-based FMA, rather than the original coordinates $\boldsymbol{X}$ the principal coordinates $\boldsymbol{P} = \boldsymbol{X}\boldsymbol{U}$ (with $\boldsymbol{U}$ as eigenvectors of $\boldsymbol{X}^t\boldsymbol{X}$, where $\boldsymbol{X}^t$ denotes the transpose of $\boldsymbol{X}$) are used instead, with the advantage that generally a smaller number of principal components $m \ll p$ can be chosen for $\boldsymbol{P}$, leading to a more stable fit, where the choice of m is data driven.

In PLS (4–6), k new regressors $\boldsymbol{T}_k$ are defined iteratively such that each coordinate is a linear combination of the original coordinates $\boldsymbol{X}$ ($\boldsymbol{T}_k = \boldsymbol{X}\boldsymbol{W}_k$) with maximal *covariance* with $\boldsymbol{f}$, while being uncorrelated to each previous coordinate in $\boldsymbol{T}_k$ (7). Subsequently, the regression problem $\boldsymbol{f} = \boldsymbol{X}\boldsymbol{W}_k\boldsymbol{\alpha}_k+\boldsymbol{\epsilon}$ is solved using $\boldsymbol{X}\boldsymbol{W}_k$ as basis. This has as an advantage that both the variance in $\boldsymbol{f}$ and $\boldsymbol{X}$ as well as the correlation between $\boldsymbol{f}$ and $\boldsymbol{X}$ is taken into account, and therefore a basis $\boldsymbol{W}_k$ is generated such that by construction includes only components of $\boldsymbol{X}$ that are correlated to $\boldsymbol{f}$ *and* have sufficient variance to contribute to $\boldsymbol{f}$. In contrast, in the PCA-based FMA the basis is selected only according to variance in $\boldsymbol{X}$.

Therefore, PLS combines the advantage of PCA-based FMA with the requirement of correlation to $\boldsymbol{f}$, thereby yielding another substantial dimension reduction ($k \leq m$) and offering a robust fit also if the number of *independent* observations is small relative to the size of the molecular system p. A priori it is not possible to estimate a proper choice for k. In fact, k serves as a regularization parameter, which has to be chosen appropriately to

maximize the predictive power, similar as m in PCA-based FMA. In practice the optimal choice is derived by cross-validation by varying k systematically until the highest correlation is obtained between $\boldsymbol{f}$ and $\boldsymbol{X}\boldsymbol{W}_k\boldsymbol{\alpha}_k$ for an independent subset of $\boldsymbol{f}$. Algorithm (8) as implemented by Denham (7) was used for the applications shown here. In this paper, this idea will be extended to an ensemble-weighted model, analogous to the ewMCM in PCA-based FMA (3). The ensemble-weighted model can be constructed from the PLS output by first converting $\boldsymbol{W}_k$ to an orthogonal basis by diagonalization of $\boldsymbol{T}_k$ and subsequently applying Eq. 12 from Hub and de Groot (3) to obtain the weights of the ewMCM. Alternatively, the ewMCM can be obtained as the scaled first column of $\boldsymbol{W}_k$.

Implementation

The PLS-based FMA has been implemented based on Helland's algorithm (8) as provided by Denham (7) and is available from the authors. Explicit details about the algorithm implementation can be found in the Supporting Material. The current implementation of the analysis tool takes coordinate trajectories in the gromacs (9) XTC format as input together with $\boldsymbol{f}$ in a generic ASCII format with two columns: One for the frame/structure identifiers and the second for the functional property of interest associated with the structure. Input coordinates should be fitted (i.e., least-squares) to a reference frame before analysis to filter out overall translation and rotation. The tool uses a preselected part of the trajectory for model building, therefore automatically allowing to use the remaining part for cross-validation. Typical computational times for ∼1000 frames of a protein of 200 amino acids (protein atoms excluding hydrogens) are in the range of a few minutes.

Results

T4 lysozyme

T4 lysozyme (T4L) is an enzyme from the bacteriophage T4 that catalyzes the hydrolysis of 1,4-beta-linkages in peptidoglycans and chitodextrins from bacterial cell walls. A prerequisite for catalysis is the correct orientation of the active site residues E11 and D20 with respect to the substrate (10). We used the distance between the C_δ of Glu11 and the C_γ of Asp20 (d_{ED}) as the functional order parameter $\boldsymbol{f}$ (Fig. 1 *A*). All backbone atoms of the protein were used as the coordinate set. Of an MD trajectory with a length of 460 ns, we used the first half for model building and the second half for cross-validation. A comparison of the results of the PLS algorithm to the PCA-based FMA is shown in Fig. 1, *B*–*G*. Correlation coefficients between model and data are shown in Fig. 1, *B* and *C* for the model building (R_m) and cross-validation (R_c) parts as a function of the number of components in the case of PLS and the number of PCA eigenvectors in the case of PCA-based FMA. The R_m for both PLS- and PCA-based FMA converge for fewer than 10 components/eigenvectors, with the PLS-based variant converging to a value closer to 1. The cross-validation R_c shows that PLS-based FMA converges at around 10 components whereas the PCA-based FMA requires a larger basis, of approx. 20 PCA eigenvectors.

In Fig. 1, *D* and *E* the overlayed MD data and PLS/PCA-based FMA model data are

shown both for the model building and cross-validation parts, using a basis of dimensionality 10 and 20 for PLS- and PCA-based FMA, respectively. These graphs show that both PLS- and PCA-based FMA provide an adequate model that cover the general features of the fluctuations in d_{ED}. The ewMCM molecular representations of these models are shown in Fig. 1, *F* and *G*. It is observed that in general terms the motions described by both models are similar, especially around the E11 and D20 amino acids.

For the PLS-based FMA we also tested if the 10 components model was accurate enough to predict the d_{ED} distance of an x-ray set of 38 T4L structures. We observe (Fig. S1, *A* in the Supporting Material) that the PLS model gives a R_c of 0.93 for these experimental structures. Then we used the T4L x-ray structures as the model building set and the MD frames as the cross-validation set. As can be seen in Fig. S1 *B* this small experimental ensemble is also able to predict correctly the MD ensemble with only 4 components. It was also tested how the reference structure (the one used for least-squares fitting the trajectory) can influence the R_m and R_c. Using reference structures with d_{ED} of 0.76, 1.01 (the one used above) and 1.24 nm showed not influence on the model quality (See Fig. S1, *C* and *D*).

Robustness

To test the consistency and robustness of the PLS- and PCA-based FMA models and the influence of the basis dimensionality, we sliced the T4L data in four equally sized parts. For each part, we built a FMA model and calculated scalar products between the MCM and ewMCM from each part (Fig. 2, *A* and *B*). In addition, cross-validation was carried out using the three parts that were not used for model building (Fig. 2, *C* and *D*). Two model dimensionalities with cross validation correlation coefficients (R_c) of approx. 0.9 were chosen for the scalar product analysis. For PLS-based FMA, we chose dimensionalities 5 and 10, whereas for PCA-based FMA we chose dimensionalities 20 and 25. In the Fig. 2, *A* and *B* we plotted the distance d_{ED} to guide the eye which part of the trajectory was used for model building. The middle and lower panels show the MCM and ewMCM scalar product matrix in a color-coded way for both PLS- and PCA-based FMA.

In general, it is observed that the overlap between the different parts is remarkably high, particularly for the MCM. Note that the FMA modes of the T4L backbone span a 1476-dimensional space (164 residues times 3 backbone atoms per residue times 3 spatial dimensions). The scalar product for two random, normalized vectors of that dimension follows a gaussian distribution with mean zero and a standard deviation of 0.026. Therefore, the probability of a scalar product of 0.5 or larger for two such random vectors is vanishingly small, at an estimated 10^{-82}. The observed scalar products therefore represent significant and substantial overlap, indicative of a robust model. The scatter in the ewMCM is by nature higher than in the MCM, as the ensemble weighting of the ewMCM introduces additional uncertainty due to non-converged variances in MD caused by incomplete sampling.

The overlap is found to be lower for a higher basis dimensionality for both PCA- and PLS-based FMA (only the ewMCM is basis independent in the case of PLS-based FMA). This indicates that a model with the lowest dimensionality that shows adequate predictive power in cross-validation should be chosen for maximal model robustness. It is interesting to note that the cross-validation correlation coefficient R_c provides a qualitative measure of MCM robustness: as can be seen in Fig. 2, trajectory parts that yield more similar models

(high scalar products) also display a larger R_c when one part is used for model training and the other for validation. This renders R_c a useful measure not only of predictive power but also of model robustness.

As an example of a highly non-linear functional property $\boldsymbol{f}$, we analyzed unfolding trajectories of the Trp-cage peptide (11) in terms of the hydrophobic solvent accessible surface (hSAS). The results are shown in the Supporting Material and Fig. S2. Surprisingly also for this non-linear case an acceptable quality model is obtained, with PLS-based FMA requiring a substantially lower dimentional basis than the PCA-based FMA.

Gating of Aquaporin channels

Yeast Aquaporin: Aqy1

Aqy1 is a tetrameric water channel of the yeast *Pichia pastoris*. The high-resolution structure revealed a closed channel, whereas functional studies indicated water channel activity (12). Together, these results therefore suggest that Aqy1 is a gated channel. Indeed, molecular dynamics simulations showed that channel opening can be reproducibly induced in response to phosphorylation of Serine 107 and by an increase of membrane pressure (mechanosensitivity) (12). By iterative, manual inspection of the trajectories it was noted that the signal was predominantly located in loop D and the lower parts of helices 4, 5 and 6. A PCA of only this region indeed identified a collective mode that correlated with channel opening events.

Here, we address the question if this collective mode can be detected unbiasedly using FMA, and tested both the PCA and PLS variants. For this study we took an MD simulation of 100 ns length of the S107D mutant of Aqpy1 and a simulation in which a lateral pressure of 10 bar in the membrane plane was applied. We used the distance between Ala190 and the center of mass of residues Pro29 and Tyr104 as $\boldsymbol{f}$, a measure of the degree of channel opening of Aqy1 (Fig. 3 *A*). For the FMA analysis we consider all the backbone atoms of each monomer and used the data of the S107D trajectory as the model training set and the lateral pressure simulation for cross-validation. In addition, we used a smaller and independent cross-validation set (30 ns) of a Aqy1 simulation where the membrane was bent toward the cytoplasmic side of the protein and also opening events were observed.

In Fig. 3 we show the comparison between the PLS- and the PCA-based FMA mode for Aqy1. Fig. 3, *B* and *C* show that the correlation between data and model in terms of both R_m and R_c is higher for the PLS as compared to the PCA-based FMA. The correlation coefficients converge to 0.9 between 10 and 20 components for the PLS-based FMA whereas the PCA-based FMA results do not yet seem to be fully converged at 100 PCA vectors. In Fig. 3, *D* and *E* it can be seen that the channel geometry data is captured adequately for the PLS-based model with 10 components and the PCA-based FMA model with 60 components. The membrane bending simulations were used as an extra and independent cross-validation set (12). Here (Fig. 3, *F* and *G*), we observe a similar trend as before: the PLS shows an acceptable model with $R_c = 0.68$ whereas PCA-based FMA shows a model with only $R_c = 0.44$, with PLS requiring a smaller basis than PCA. The ewMCM representation of the PLS-based FMA fluctuations is shown in Fig. 3 *H*. The PLS- and the PCA-based FMA versions of the ewMCM have a scalar product of 0.99, and show collective motions in the protein. Backbone motions involve primarily loop D and the lower halves of helices 1, 3, 4,

and 6 which are coupled to the local opening of the pore. Hence, the ewMCM provides an explanation of how the gating residues, that are not in direct contact with the membrane, are affected by changes in the membrane, either induced by an applied lateral pressure, or by membrane bending. The signal seems to be transmitted from helix 1 and 6, that are in direct contact with the membrane to the lower parts of helices 3, 4 and 5, and loop D, that line the water pore.

We compared the first PCA eigenvector of the lower parts of helix 4, 5, 6 and loop D as described in (12) with the PLS-based FMA of the full backbone of Aqy1. The scalar product of both modes in this subset of atoms is 0.66 which implies a high degree of similarity of the motions. Together, in this case, these results show that unbiased PLS-based FMA analysis gives similar modes compared with a selective/iterative PCA analysis as described in (12) (see Fig. S3).

Human aquaporin-1: hAQP1

hAQP1 is a tetrameric water channel ubiquitously expressed in the cell membranes. The x-ray structure of its high-identity bovine homolog (13) shows two constrictions for the water conduction: The NPA signature motif and the aromatic/arginine (ar/R) site, the later formed by R195, H180 and F56. Molecular dynamics simulations showed that channel opening and closing could be induced in response to voltage changes (14) in the range of –1.5 to 1.5 V. Those analyses showed a correlation between the permeability coefficients (p_f) and the membrane potential, with the channel more open at positive potentials. It was also reported that the flipping of the R195 side chain is involved in the open-close transitions.

Here, we address the question if we can find a global structural model that is able to explain the changes in the functional property p_f using PLS-based FMA. For this purpose we took 22 MD simulations of 60 ns length calculated in a double membrane setup at $\pm$ 1.5 V (14). We calculated the single-channel permeabilities p_f from the collective diffusion model proposed by Zhu *et al.* (15) at a single monomer basis. Because p_f is a property that does not depend on a single structure, p_f values were calculated using the last 50 ns of each trajectory, using 5 ns windows. For the FMA analysis we used the average structures of the monomer atoms (excluding the hydrogens) of the same time windows used for the p_f calculation. Since we had 8 monomers in total in the double membrane setup, we used 6 of them for model building and 2 for cross-validation.

In Fig. 4, *A* we show the correlation between data and model in terms of both R_m and R_c for the PLS-based FMA. The correlation coefficients of the training part converged to 0.9 around 30 components. The R_c values converged to 0.6 with the same number of components. Note in Fig. 4, *B* that the p_f signal intrinsically suffers from a low signal to noise ratio. So, the favorable correlation in the cross-validation stage is remarkable. A ewMCM representation of the PLS-based FMA fluctuations and extremes are shown in Fig. 4 *C* and *D*. In terms of fluctuations Fig. 4 *C* shows the ewMCM changes in the loops and in the backbone of the extracellular half of the protein, mainly around R195. The extremes representation of the ewMCM in Fig. 4 *D* shows that the displacement of R195 side chain that changes its distance to H180, as previously suggested from visual inspection (14). In addition, we noted the displacement of N127, F212, I211 and W210 which seems to move correlatedly with R195. Interestingly, the R195V mutant in rat AQP1 does not change the water permeability

but allows urea, glycerol, amonia and protons to pass (16). Simlilarly, R195S shows proton and cation permeability and a higher osmotic water permeability (17).

Conformational transitions of CLC-ec1

CLC comprises a family of transport proteins that function as chloride channels or proton/chloride exchangers (18). CLCs share a similar fold (18 α-helices, labeled from A to R) and dimeric architecture as it observed in x-ray structures from bacteria to eukaryotes (19, 20). Mutation and electrophysiology studies have identified a glutamate in the selectivity filter which is essential for the exchange mechanism and for gating in the channels counterpart (21–24). In CLC-ec1 from *Escherichia coli* this conserved glutamate (E148) resides in the selectivity filter, between the extracellular and intracellular vestibules of the protein (Fig. 5 *A*). Close to E148 two chlorides can be found in the wild-type, which define the central and internal binding sites for anions (S_{cen} and S_{int}). It has been shown by x ray (21), MD (25) and Metadynamics (26) that E148 shows an intrinsic flexibility which may play a role in transport mechanisms.

By using MD, PLS-based FMA and electrostatic calculations we show that the intrinsic flexibility of E148 also depends on the chloride occupation and these changes are related to local and global changes in the CLC structure. We simulated wild-type CLC-ec1 (unprotonated state of E148) inserted in a pre-equilibrated 1-palmitoyl-2-oleoyl-*sn*-glycero-3-phosphoethanolamine (POPE) membrane patch. We ran six different CLC-ec1 simulations for which we changed and enforced the chloride occupation of S_{cen} and S_{int}. The ion configurations were: *No ions*; S_{int} restrained; S_{cen} restrained; S_{int} and S_{cen} restrained; S_{int} restrained and S_{cen} free; free S_{int} and S_{cen}. In restraining the ions we used a force constant of 1000 [kJ mol^{-1} nm^{-2}]. The simulations were ran 100 ns each.

The simulations were stable and showed no spontaneous chloride translocations events. In contrast, E148 showed spontaneous flexibility in the simulations. The flexibility corresponded to transitions of the glutamate from the α-helical to the β-sheet zone in a Ramachandran space. To quantify the E148 variability we calculated the Ψ dihedral angle distribution of each simulation. In the Fig. 5 *B* we observe the Ψ histograms which show three main peaks around -75° (α), 25° (I for intermediate) and 120° (β). Interestingly, the Ψ angle distributions clearly correlate with the single anion occupation or absence of anions in the selectivity filter. When the protein is occupied at S_{cen} the main E148 conformation is α; when occupied at S_{int} the main conformation is β; and when the protein lacks anions at the selectivity filter sites, E148 adopts mainly the intermediate conformation. The doubly occupied monomers showed different proportions of the three Ψ peaks (Fig. S4). Structurally, the conformations of E148 imply a change in the backbone atoms of the glutamate and glycine of the highly conserved sequence *GR**EG**P* (19) which flank the central site S_{cen}. They change from orienting the amide nitrogen of G149 toward S_{cen} in the α conformation, to orient the carbonyl group of E148 in β. In the intermediate state (I) the peptidic bond between these amino acids is parallel to S_{cen}.

We used PLS-based FMA to understand the structural changes of the CLC-ec1 protein related with changes in the E148 Ψ angle. For that we calculated FMA in a monomer basis using the protein excluding the hydrogens atoms. We constructed the FMA models using 75 % of the data and the remaining 25 % for cross-validation. In Fig. 5 *C* we observe

that R_m converges to values close to 1 after 40 PLS components, whereas R_c reaches values above 0.8 for 10 or more PLS components and above 0.9 after 40 components. An overlay of the original data set, the model and the cross-validation for the model built using 40 PLS components is shown in Fig. 5 *D*, showing that the FMA model recovers most of the features of the original Ψ data. The structural changes observed in the ewMCM model show delocalized fluctuations in the whole monomer, (Fig. 5 *E*) mainly in the loops B-C, F-G, I-J, K-L, O-P and helices Q and R. At E148 the ewMCM includes the transition of its backbone atoms from -70° to 90° in the Ψ angle.

As a next step we calculated pore radii profiles (using *Mole* (27)) and the electrostatic potential (using APBS (28)) along the ewMCM. We used the position of E148 as starting point for *Mole* paths searches that connected the central site (S_{cen} around 44 Å in the z coordinate) with the extracellular side or with the intracellular vestibule (by S_{int} around 38 Å) through the chloride path (29). Fig. 5 *F* (upper panel) shows the radii and the electrostatic profile projected onto the z coordinate for interpolated frames along the ewMCM. We selected frames corresponding to α, I and β conformations of E148. The radii profiles show similar trends for the three frames. i.e., all pore profiles show a constriction below 2 Å radius between 37 and 53 Å. Within this zone they also show a 2 Å peak at the location of S_{cen}. Frames I and β also show an increase of the radii between S_{int} to 30 Å, in the intracellular part of the chloride path. The electrostatic potentials (Fig. 5 *F* lower panel) along these paths was multiplied by -1 to display the attractive potential for anions (positive) as wells. Profiles are more attractive to anions in the intracellular side and slightly repulsive toward the extracellular side. The most dramatic changes in the potentials occur in the zone between 38 to 48 Å. We observe a first well around 39 Å (S_{int}) which increases systematically along the ewMCM (α to β from -22 to -10 k_BT/e). Similarly, the well around 44 Å (S_{cen}) changes form -28 to 4 k_BT/e. From the S_{cen} site to the extracellular site we found some discontinuities in the paths, which correspond to the most constricted zones of the channel (radii below 1 Å). We speculate that these changes in electrostatic potential and radii are inherent to the occupation of chlorides in CLC. These changes may modulate the relative affinity and accessibility of the sites in the transport cycle of these proteins, therefore directly linking local changes in E148 to global changes in the CLC-ec1 structure. In a transport context, the changes along the Ψ angle show opening of the intracellular chloride path. These changes include the motion of the helices Q,R and Y455 which have been indicated (30) as part of an internal gate in CLC-ec1. Also, the changes show how the anion occupation can tweak the electrostatic potential at S_{cen} (31), where anions and protons can go through.

Discussion

The applications presented in the results section demonstrate that PLS-based FMA provides a general method to identify a hidden relation between coordinates and a functional order parameter $\boldsymbol{f}$ of interest. In the current implementation only a unidimensional $\boldsymbol{f}$ is allowed. It yields a linear model in the form of a collective mode of dynamics that optimizes the covariance with the observed data. This collective mode allows a direct interpretation of the relation between the functional order parameter and the underlying protein mechanics. It also allowd to make hypothesis about the relevant amino acids contributing the most to the

funtional property.

Due to the inherent overfitting risk encountered in fitting high-dimensional data sets, cross-validation with independent data is a mandatory step to assess model quality. In all investigated cases, a satisfactory correlation coefficient between model and data was obtained for cross-validation subsets (R_c) of the data that did not substantially deviate from the training subsets (R_m). Increasing the dimensionality of the basis leads to an ever increasing R_m, but to a R_c that goes through a maximum and then deteriorates due to overfitting.

PLS-based FMA models derived from independent trajectories were found to be remarkably similar, indicating that the models, represented as MCM, are a robust representation of the relation between $\boldsymbol{f}$ and the atomic coordinates (Fig. 2). Naturally, the ewMCM scatters more for different independent trajectories, as the limited sampling in each (sub)trajectory will affect the ensemble weighting. This effect is analogous to the observation that the eigenvalues along individual PCA modes converge slowly in MD (32). It is interesting to note that the cross-validation correlation coefficient R_c provides a qualitative measure of MCM robustness: as can be seen in Fig. 2, trajectory parts that yield more similar models (high scalar products) also display a larger R_c when one part is used for model training and the other for validation.

The robustness assessment shown in Fig. 2 also indicates that the robustness decreases when increasing the dimensionality of the basis, even for a dimensionality where R_c does not yet indicate overfitting. This is likely due to coordinates with relatively little variance that on the one hand aid to marginally improve the model (as probed by R_c), but on the other hand deteriorate model robustness by including additional dimensions. For most pratical purposes, a minimal basis dimensionality with adequate R_c should therefore be prefered.

A prerequisite for the application of FMA is the availability of a suitable functional order parameter $\boldsymbol{f}$. This poses a limitation for cases where a unique parameter of functional interest cannot be uniquely defined.

The current PLS-based implementation is restricted to linear correlations. As shown before, FMA can be extended to general correlations based on mutual information (MI) (3). An extention to a MI based implementation is considered for the future.

Conclusions

We have introduced a versatile and general approach to relate an external order parameter $\boldsymbol{f}$ to a collective mode of internal dynamics. The partial least squares algorithm proved to yield robust solutions to the underlying multidimensional regression problem with minimal and controlable overfitting risk. The aquaporins and CLC-ec1 examples of the PLS-based functional mode analysis illustrate that the approach succesfully captures the relation between internal protein dynamics and different functional order parameters of interest. For the Aqy1 case a putative coupling between the membrane-facing surface and the inner water pore was identified, for hAQP1 the osmotic permeability p_f was shown to relate mostly with changes around ar/R region, and for CLC-ec1 the local mobility of the gating residue Glu-148 was found to be coupled to a collective mode that may modulate the chloride ion binding affinity in pore locations remote from the gating residue. These examples illustrate that PLS-based FMA can be successfully used to study functional mechanisms by detecting

collective modes of dynamics that are most related to the fluctuation of a functional property of interest. In addition, such modes can be explored dynamically for additional functional states using techniques like essential dynamics sampling (33) or conformational flooding (34).

Acknowledgements

We thank Camilo Aponte-Santamaría for the Aqy1 MD trajectories. T.K. and A.M. are grateful of N. Krämer and G. Blanchard for helpful comments. We thank Oliver Beckstein for carefully reading the manuscript. We gratefully acknowledge the DFG for funding, via the collaborative research grant SFB803, FOR916 as well as support through a Marie-Curie Intra-European fellowship within the 7$^{\text{th}}$ European Community framework programme. T.K. acknowledges the support of the DFG as part of the institutional strategy of the Georg-August-University Göttingen.

Supporting Citations

References (35–39) appear in the Supporting Material.

References

1. Garcia, A. E., 1992. Large-amplitude nonlinear motions in proteins. *Phys. Rev. Lett.* 68:2696–2699.

2. Amadei, A., A. B. M. Linssen, and H. J. C. Berendsen, 1993. Essential dynamics of proteins. *Proteins: Struct. Funct. Genet.* 17:412–425.

3. Hub, J. S., and B. L. de Groot, 2009. Detection of functional modes in protein dynamics. *PLoS Comput. Biol.* 5:e1000480.

4. Hold, H. O. A., 1966. Nonlinear estimation by iterative least squares procedures. *In* F. N. David, editor, Research papers in statistics: Festschrift for J. Neyman, Wiley, New York, 411–444.

5. Hold, H. O. A., 1973. Nonlinear iterative partial least squares (NIPALS) modelling: some current developments. *In* P. Krishnaiah, editor, Multivariate analysis III, Academic Press, New York, 383–407.

6. Hold, H. O. A., 1982. Soft modeling: the basic design and some extensions. *In* K. G. Jöreskog, and H. O. A. Wold, editors, Systems under indirect observation: causality, structure, prediction, North-Holland, Amsterdam.

7. Denham, M. C., 1995. Implementing partial least squares. *Stat. Comput.* 5:191–202.

8. Helland, I. S., 1988. On the structure of partial least squares regression. *Commun. Stat. Simulat.* 17:581–607.

9. Van der Spoel, D., E. Lindahl, B. Hess, G. Groenhof, A. E. Mark, and H. J. C. Berendsen, 2005. GROMACS: fast, flexible and free. *J. Comp. Chem.* 26:701–1719.

10. Phillips, D. C., 1967. The hen egg-white lysozyme molecule. *P. Natl. Acad. Sci. USA* 57:483–495.

11. Neidigh, J. W., R. M. Fesinmeyer, and N. H. Andersen, 2002. Designing a 20-residue protein. *Nat. Struct. Biol.* 9:425–430.

12. Fischer, G., U. Kosinska-Eriksson, C. Aponte-Santamaría, M. Palmgren, C. Geijer, K. Hedfalk, S. Hohmann, B. L. de Groot, R. Neutze, and K. Lindkvist-Petersson, 2009. Crystal structure of a yeast aquaporin at 1.15 Å reveals a novel gating mechanism. *PLoS Biol.* 7:e1000130.

13. Sui, H., B. Han, J. Lee, P. Walian, and B. Jap, 2001. Structural basis of water-specific transport through the AQP1 water channel. *Nature* 414:872–878.

14. Hub, J. S., C. Aponte-Santamaria, H. Grubmueller, and B. L. de Groot, 2010. Voltage-Regulated Water Flux through Aquaporin Channels In Silico. *Biophys. J.* 99:L97–L99.

15. Zhu, F., E. Tajkhorshid, and K. Schulten, 2004. Collective diffusion model for water permeation through microscopic channels. *Phys. Rev. Lett.* 93:224501.

16. Beitz, E., B. Wu, L. Holm, J. Schultz, and T. Zeuthen, 2006. Point mutations in the aromatic/arginine region in aquaporin 1 allow passage of urea, glycerol, ammonia, and protons. *P. Natl. Acad. Sci. USA* 103:269–274.

17. Li, H., H. Chen, C. Steinbronn, B. Wu, E. Beitz, T. Zeuthen, and G. A. Voth, 2011. Enhancement of Proton Conductance by Mutations of the Selectivity Filter of Aquaporin-1. *J. Mol. Biol.* 407:607–620.

18. Chen, T.-Y., 2005. Structure and function of CLC channels. *Annu. Rev. Physiol.* 67:809–839.

19. Dutzler, R., E. B. Campbell, M. Cadene, B. T. Chait, and R. MacKinnon, 2002. X-ray structure of a CLC chloride channel at 3.0 Å reveals the molecular basis of anion selectivity. *Nature* 415:287–294.

20. Feng, L., E. B. Campbell, Y. Hsiung, and R. MacKinnon, 2010. Structure of a eukaryotic CLC transporter defines an intermediate state in the transport cycle. *Science* 330:635–641.

21. Dutzler, R., E. B. Campbell, and R. MacKinnon, 2003. Gating the selectivity filter in CLC chloride channels. *Science* 300:108–112.

22. Estévez, R., B. C. Schroeder, A. Accardi, T. J. Jentsch, and M. Pusch, 2003. Conservation of chloride channel structure revealed by an inhibitor binding site in CLC-1. *Neuron* 38:47–59.

23. Yusef, Y. R., L. Zúñiga, M. Catalán, M. I. Niemeyer, L. P. Cid, and F. V. Sepúlveda, 2006. Removal of gating in voltage-dependent CLC-2 chloride channel by point mutations affecting the pore and C-terminus CBS-2 domain. *J. Physiol-London* 572:173–181.

24. Zdebik, A. A., G. Zifarelli, E.-Y. Bergsdorf, P. Soliani, O. Scheel, T. J. Jentsch, and M. Pusch, 2008. Determinants of anion-proton coupling in mammalian endosomal CLC proteins. *J. Biol. Chem.* 283:4219–4227.

25. Bostick, D. L., and M. L. Berkowitz, 2004. Exterior site occupancy infers chloride-induced proton gating in a prokaryotic homolog of the CLC chloride channel. *Biophys. J.* 87:1686–1696.

26. Gervasio, F. L., M. Parrinello, M. Ceccarelli, and M. L. Klein, 2006. Exploring the gating mechanism in the CLC chloride channel via metadynamics. *J. Mol. Biol.* 361:390–398.

27. Petřek, M., P. Košinová, J. Koča, and M. Otyepka, 2007. MOLE: a Voronoi diagram-based explorer of molecular channels, pores, and tunnels. *Structure* 15:1357–1363.

28. Baker, N. A., D. Sept, S. Joseph, M. J. Holst, and J. A. McCammon, 2001. Electrostatics of nanosystems: application to microtubules and the ribosome. *P. Natl. Acad. Sci. USA* 98:10037–10041.

29. Accardi, A., M. Walden, W. Nguitragool, H. Jayaram, C. Williams, and C. Miller, 2005. Separate ion pathways in a Cl^-/H^+ exchanger. *J. Gen. Phys.* 126:563–570.

30. Jayaram, H., A. Accardi, F. Wu, C. Williams, and C. Miller, 2008. Ion permeation through a Cl^--selective channel designed from a CLC Cl^-/H^+ exchanger. *P. Natl. Acad. Sci. USA* 105:11194–11199.

31. Zhang, Y., and G. Voth, 2011. The coupled proton transport in the CLC-ec1 Cl^-/H^+ Antiporter. *Biophys. J.* 101:47–49.

32. de Groot, B. L., D. M. F. van Aalten, A. Amadei, and H. J. C. Berendsen, 1996. The consistency of large concerted motions in proteins in molecular dynamics simulations. *Biophys. J.* 71:1707–1713.

33. Amadei, A., A. B. M. Linssen, B. L. de Groot, D. M. F. van Aalten, and H. J. C. Berendsen, 1996. An efficient method for sampling the essential subspace of proteins. *J. Biom. Str. Dyn.* 13:615–626.

34. Grubmüller, H., 1995. Predicting slow structural transitions in macromolecular systems: conformational flooding. *Phys. Rev. E.* 52:2893–2906.

35. Krämer, N., A.-L. Boulesteix, and G. Tutz, 2008. Penalized partial least squares with applications to B-spline transformations and functional data. *Chemometr. Intell. Lab.* 94:60–69.

36. Manne, R., 1987. Analysis of two partial least squares algorithms for multivariate calibration. *Chemometr. Intell. Lab.* 2:187–197.

37. Bühlmann, P., and B. Yu, 2003. Boosting with the L2 loss: regression and classification. *J. Am. Stat. Assoc.* 98:324–339.

38. Bissantz, N., T. Hohage, A. Munk, and F. Ruymgaart, 2007. Convergence rates of general regularization methods for statistical inverse problems and applications. *SIAM J. Numer. Anal.* 45:2610–2636.

39. Blanchard, G., and N. Krämer, 2010. Optimal learning rates for kernel conjugate gradient regression. *In* Adv. Neur. In. volume 23, 226–234.

Figure Legends

Figure 1.

Comparison of PLS- and PCA-based FMA methods for Glu11-Asp20 distance d_{ED} (A) of T4 Lysozyme (T4L). (B/C) Pearson correlation coefficients between data and model for PLS- and PCA-based FMA as function of the number of PLS components or PCA vectors calculated for the model training subset (*black*, R_m) and the cross-validation subset (*red*, R_c). (D/E) Overlay of data and model for the calculated distances d_{ED} as function of time. The *black lines* correspond to the MD data, the *green* to the model training subset and *red* to the model cross-validation subset. The models were calculated using 10 components for PLS-based FMA (D) and 20 PCA eigenvectors for PCA-based FMA (E). (F/G) Backbone representation of the ensemble-weighted MCM (ewMCM) contributing to the change in the distance d_{ED}. The *red-white-blue* color-scaled structures represent the interpolation between the extremes of the ewMCMs. The PLS- and PCA-based FMA models used to plot the molecular representations have the same number of components or PCA vectors as the panels D and E.

Figure 2.

Scalar product analysis of PLS- and PCA-based FMA MCM and ewMCM models derived from different trajectory parts applied to the distance d_{ED} of T4 Lysozyme. (A) and (B) The upper panels show the T4L distance d_{ED} as function of time. The x axis is divided in four equally spaced sub-trajectories. The mid and lower panels correspond to color-coded matrices of the scalar products of MCM and ewMCM vectors, respectively. Each (ew)MCM vector was calculated using one fourth of the T4L trajectory, as indicated in the upper panel. All the scalar product combinations were calculated for 5 and 10 components for PLS-based FMA, and 20 and 25 eigenvectors in the case of PCA-based FMA. Panels C and D show cross-validation correlation coefficients R_c as a function of the basis dimensionality. Indices of the form i_j mean that the model was constructed for fragment i and cross-validated with fragment j.

Figure 3.

Comparison of PLS- and PCA-based FMA for the degree of channel opening of yeast Aquaporin (Aqy1). (A) The degree of opening was defined as the distance between Ala190 and

the center of mass of residues Pro29 and Tyr104. Each monomer was considered an independent channel and therefore the four monomer trajectories were concatenated. The 100 ns simulation of the S107D mutant used for model training, therefore represent the first 400 ns of the concatenated trajectory, and the applied lateral pressure simulations the time window from 400–800 ns. (*B/C*) Pearson correlation coefficients between data and model for PLS/PCA-based FMA as a function of the number of PLS components/PCA vectors calculated for the model training subset (*black*, R_m) and the cross-validation subset (*red*, R_c). (*D/E*) Overlay of data and model for the calculated channel opening distance as function of time. The *black lines* correspond to the MD data, the *green* to the model training subset and *red* to the cross-validation subset. (*F/G*) 120 ns of a membrane bending simulation were used as an extra cross-validation sets (*violet line*). The models were calculated using 10 components for PLS (*D*) and 60 PCA vectors for PCA-based FMA (*E*). (*H*) Side and bottom view backbone representations of the PLS-based FMA ewMCM contributing to the change in the channel gating distance. The color-scale (*blue-green-red*) and the line thickness represents the RMSF of the ewMCMs. S107, Y104, P29 and A190 backbone atoms are shown as spheres. The model used for PLS- based FMA ewMCM representations has the same number of components as the model in panel *D*.

Figure 4.

PLS-based FMA for single-channel water permeability (p_f) of human Aquaporin 1 (hAQP1) in the presence of a transmembrane voltage. The p_f and protein average structures were calculated within 5 ns trajectory windows and each monomer was considered an independent channel. The simulations of 6 monomers were used for model training, and 2 independent monomers for cross-validation. (*A*) Pearson correlation coefficients between data and model for PLS-based FMA as a function of the number of PLS components was calculated for the model training subset (*black*, R_m) and the cross-validation subset (*red*, R_c). (*B*) Overlay of data and model for the calculated p_f values using 30 components (*arrow* in panel *A*) as a function of time. The *black lines* correspond to the MD data, the *green* to the model training subset and *red* to the cross-validation subset. (*C*) Backbone and stick representations of the amino acids in the ewMCM mode contributing to the change in p_f. The color-scale (*blue-green-red*) and the thickness of the lines represent the root mean-square fluctuation of the ewMCM. *Red* and thicker sticks means amino acids with higher RMSF. (*D*) Cartoon and overlay representation of the helices and amino acids contributing to the ewMCM. The color-scale (*red-green-blue*) represent the conformations, associated with low (*red*), intermediate (*green*) and high (*red*) estimated p_f values. The side chain motion of R195 asociated with the low and high p_f values is highlighed by the *curved arrow*. The locations of R195, H180, N127, F212, I211, W210 and K36 amino acids are indicated in the panels (*C*) and (*D*) by sticks of their corresponding RMSF or p_f conformation color. The model used for PLS-based FMA ewMCM representations have the same number of components as the model in panel *B*.

Figure 5.

PLS-based FMA analysis for the Ψ angle of the gating glutamate E148 of *E. coli* CLC protein. (*A*) Cartoon representation of CLC-ec1 dimer. E148 is shown in ball and stick representation. Chlorides occupying the crystallographic binding sites S_{cen} and S_{int} are shown as green spheres. (*B*) Distribution of the E148 Ψ angle for single occupied monomers at S_{cen} and S_{int}, or for empty monomers (*No-ions*). S_{cen}-occupied monomers mostly populate the α-helical conformation (*blue*), S_{int}-occupied populate the β-sheet conformation (*red*) and in the empty monomers E148 populate an intermediate (*I, green*) conformation, between α and β. (*C*) Pearson correlation coefficients between data and model for PLS-based FMA as a function of the number of PLS components calculated for the model training subset (*black*, R_m) and the cross-validation subset (*red*, R_c). (*D*) Overlay of data and model for the calculated CLC-ec1 angle Ψ_{E148} as function of time. The *black lines* correspond to the MD data, the *green* to the model training subset and *red* to the cross-validation subset. (*E*) Backbone representations of the PLS-based FMA ewMCM contributing to the change in the Ψ_{E148} angle. (*left*) The color-scale (*blue-green-red*) and the thickness of the lines represent the root mean-square fluctuation of the ewMCMs. (*right*) Cartoon and overlay representation of the helices and loops of CLC-ec1, *Blue* represent the α conformation, *green* the intermediate and *red* the β conformation. The locations of E148, S107, and Y445 are indicated in both representations by spheres or sticks. The model used for PLS-based FMA ewMCM representations have the same number of components as the model in panel *D*. (*F*) Path radius and electrostatic potential calculated along the ewMCM. The conformations correspond to α, I and β for Ψ_{E148}. The potential was multiplied by -1 to visualize as wells the attractive potential on negative particles (like chlorides). *Arrows* indicate the location of the central and internal chloride sites.

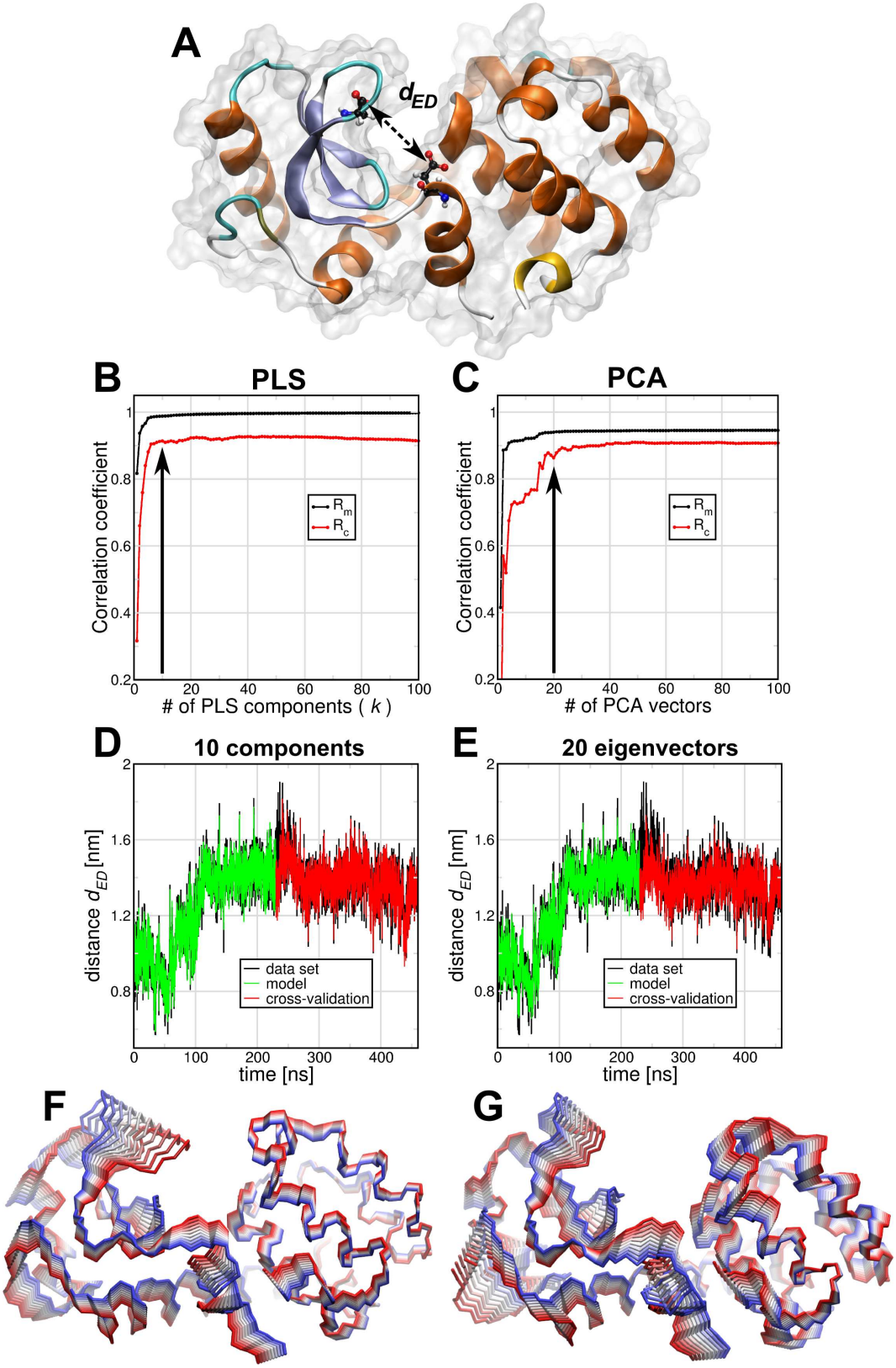
A
d_{ED}
B
PLS
Correlation coefficient
# of PLS components (k)
R_m
R_c
C
PCA
Correlation coefficient
# of PCA vectors
D
10 components
distance d_{ED} [nm]
time [ns]
data set
model
cross-validation
E
20 eigenvectors
distance d_{ED} [nm]
time [ns]
F
G

Figure 1:

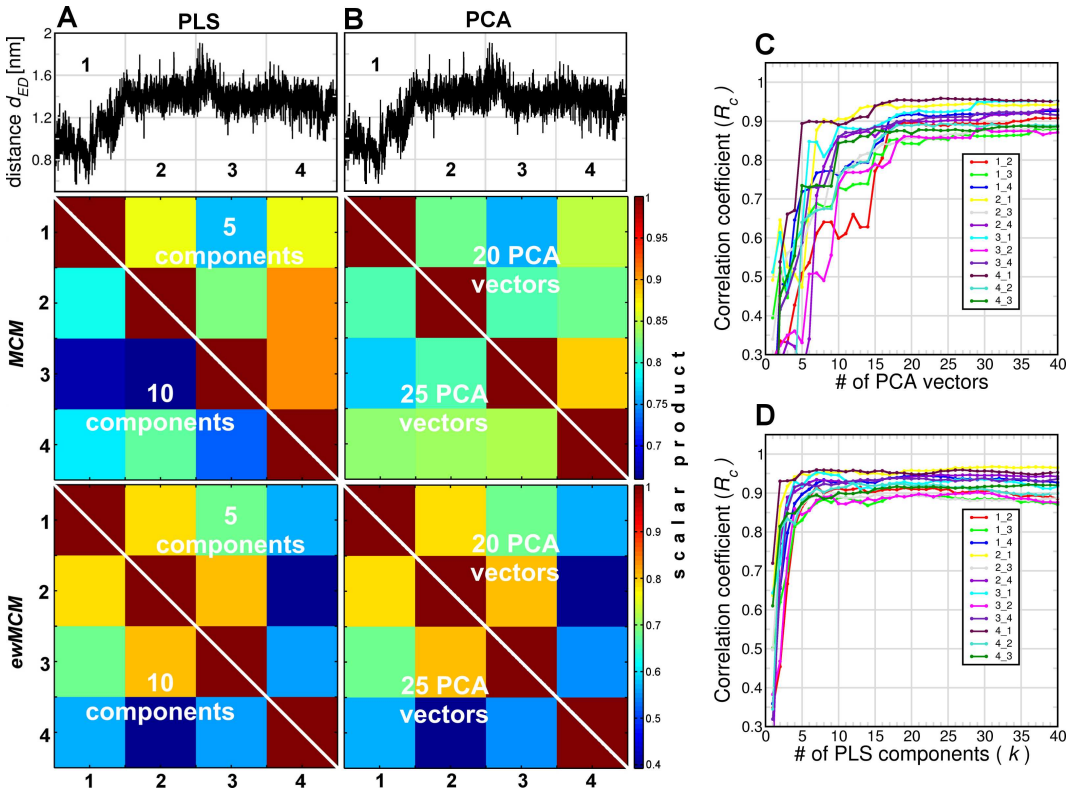
A
PLS
B
PCA
C
D
distance d_{ED} [nm]
MCM
ewMCM
5 components
10 components
20 PCA vectors
25 PCA vectors
scalar product
Correlation coefficient (R_c)
# of PCA vectors
# of PLS components (k)

Figure 2:

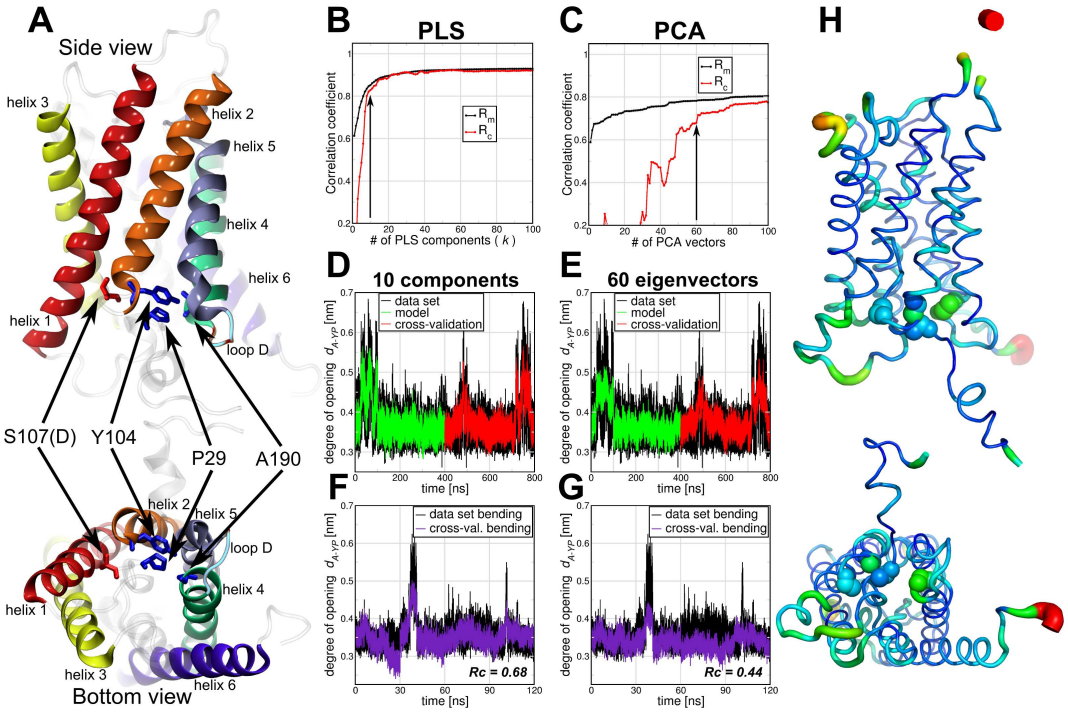
A
Side view
helix 3
helix 2
helix 5
helix 4
helix 6
helix 1
loop D
S107(D)
Y104
P29
A190
helix 2
helix 5
loop D
helix 4
helix 1
helix 3
helix 6
Bottom view
B
PLS
Correlation coefficient
# of PLS components (k)
C
PCA
Correlation coefficient
# of PCA vectors
D
10 components
data set
model
cross-validation
degree of opening
time [ns]
E
60 eigenvectors
data set
model
cross-validation
degree of opening
time [ns]
F
data set bending
cross-val. bending
Rc = 0.68
degree of opening
time [ns]
G
data set bending
cross-val. bending
Rc = 0.44
degree of opening
time [ns]
H

Figure 3:

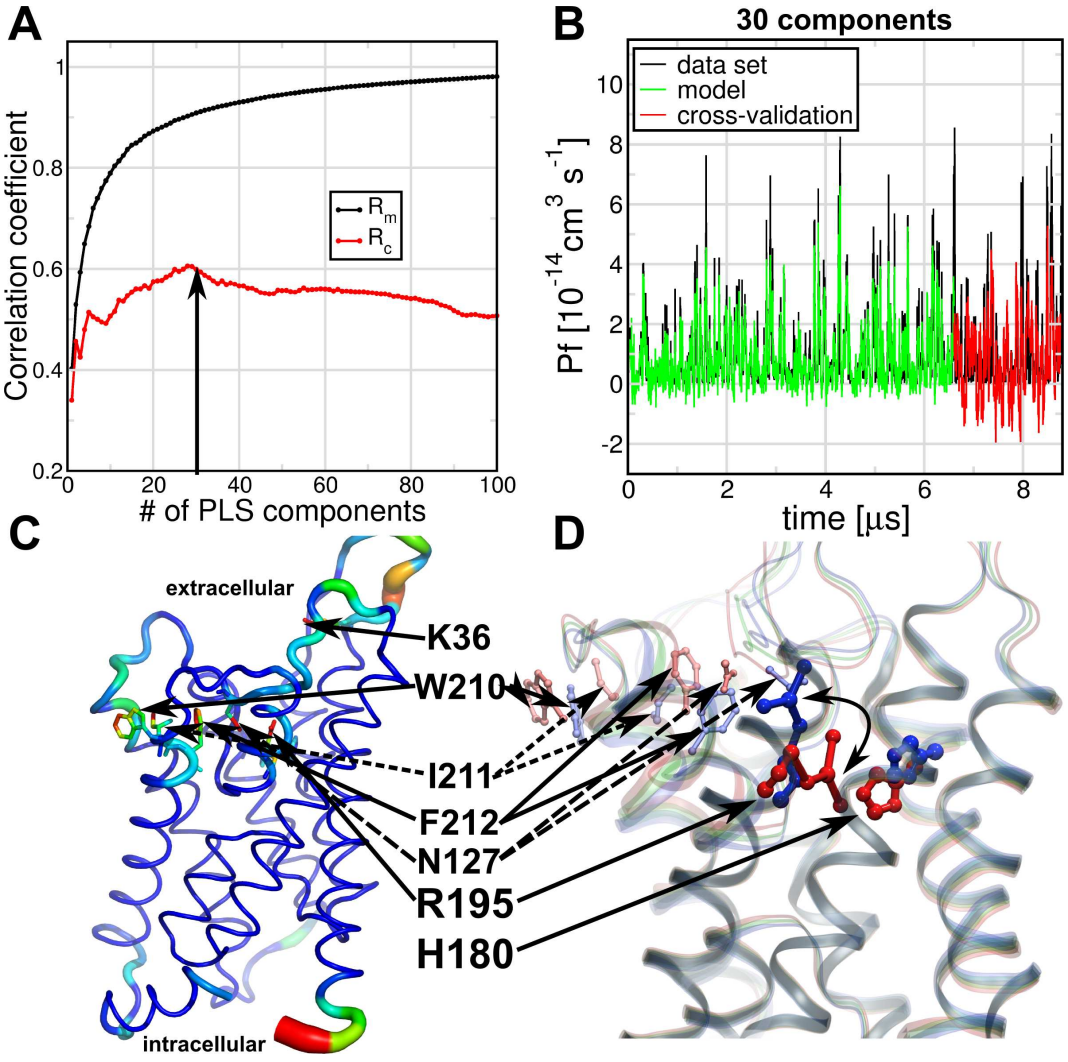
A
Correlation coefficient
# of PLS components
R_m
R_c
B
30 components
data set
model
cross-validation
Pf [10^-14 cm^3 s^-1]
time [μs]
C
extracellular
intracellular
K36
W210
I211
F212
N127
R195
H180
D

Figure 4:

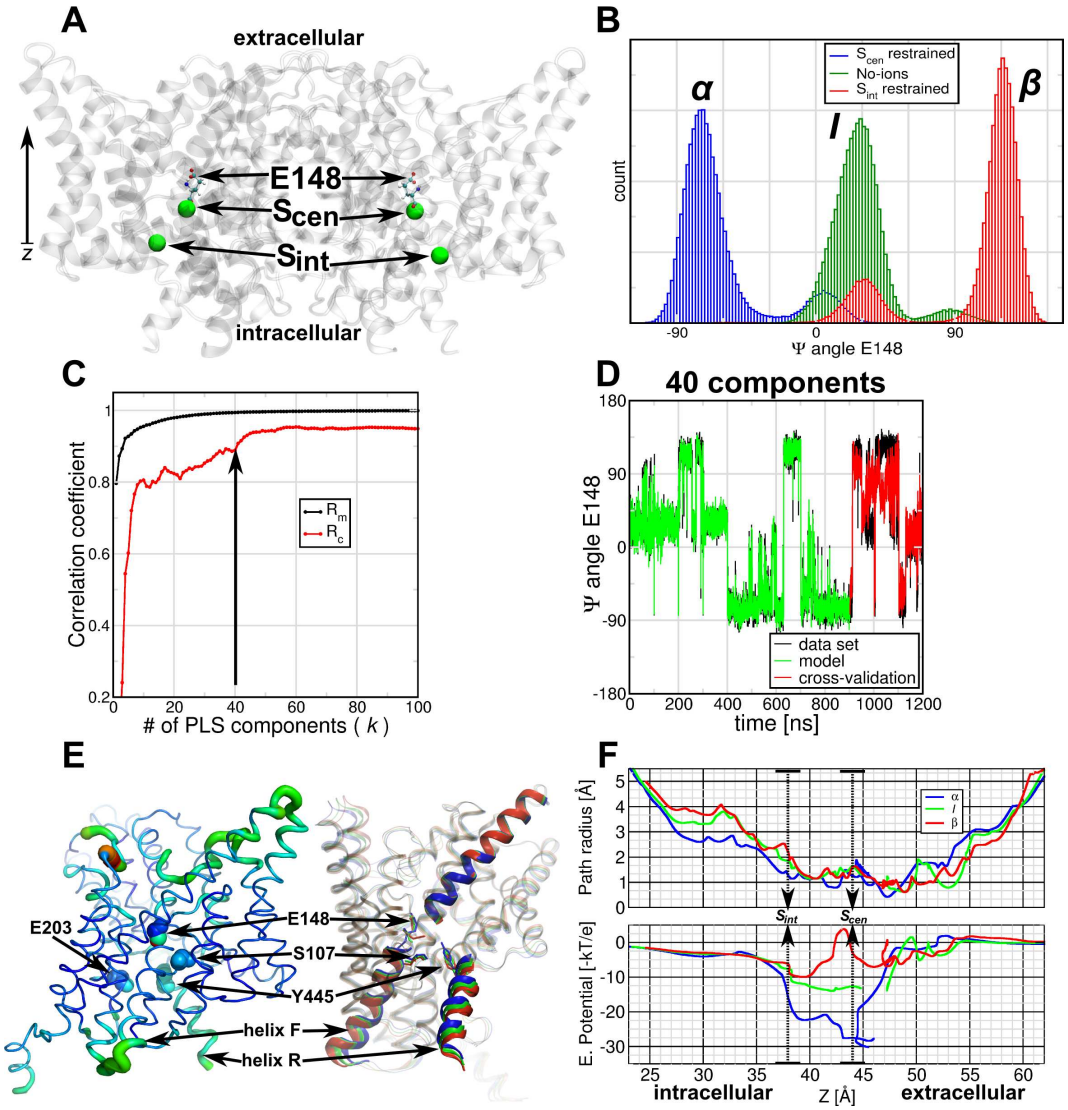
A
extracellular
E148
Scen
Sint
z
intracellular
B
α
l
β
Scen restrained
No-ions
Sint restrained
count
-90
0
90
Ψ angle E148
C
Correlation coefficient
Rm
Rc
# of PLS components (k)
D
40 components
Ψ angle E148
data set
model
cross-validation
time [ns]
E
E203
E148
S107
Y445
helix F
helix R
F
Path radius [Å]
α
l
β
Sint
Scen
E. Potential [-kT/e]
intracellular
Z [Å]
extracellular

Figure 5: