Effect of ligand binding on the intraminimum dynamics of proteins

Effect of Ligand Binding on the Intraminimum Dynamics of Proteins

BURAK ALAKENT,[1,2*] SENA BASKAN,[3,4] PEMRA DORUKER[1,3,5*]
[1]*Department of Chemical Engineering, Bogazici University, Bebek, Istanbul 34342, Turkey*
[2]*Department of Chemical Engineering, Yeditepe University, Kadikoy, Istanbul 34755, Turkey*
[3]*Polymer Research Center, Bogazici University, Bebek, Istanbul 34342, Turkey*
[4]*Computational Science and Engineering Program, Bogazici University, Bebek, Istanbul 34342, Turkey*
[5]*Feza Gursey Institute, Cengelkoy, Istanbul 34684, Turkey*

Received 23 September 2009; Revised 2 June 2010; Accepted 23 June 2010
DOI 10.1002/jcc.21636
Published online 20 August 2010 in Wiley Online Library (wileyonlinelibrary.com).

Abstract: Effects of ligand binding on protein dynamics are studied via molecular dynamics (MD) simulations on two different enzymes, dihydrofolate reductase (DHFR) and triosephosphate isomerase (TIM), in their unliganded (free) and liganded states. Domain motions in MD trajectories are analyzed by collectivities and rotation angles along the principal components (PCs). DHFR in the free state has well-defined domain rotations, whereas rotations are slightly damped in the binary complex with nicotinamide adenine dinucleotide phosphate (NADPH), and remarkably distorted in the presence of $NADP^+$, showing that $NADP^+$ is solely responsible for the loss of correlation of the domains in DHFR. Although mean square fluctuations of MD simulations in the same PC subspaces are similar for different ligation states, linear stochastic time series models show that backbone flexibility along the first five PCs is decreased upon NADPH and $NADP^+$ binding in subpicosecond scale. This shows that mobility of the protein along the PCs is closely related with intraminimum dynamics, and alterations in ligation states may change the intraminimum dynamics significantly. Low vibrational frequencies of the alpha-carbon atoms of DHFR are determined from the time series models of a larger number of low indexed PCs, and it is found that number of modes in the lowest frequencies is reduced upon ligand binding. A similar result is obtained for TIM in the unliganded and dihydroxyacetone phosphate bound states. We suggest that stochastic time series modeling is a promising method to be used in determining subtle perturbations in protein dynamics.

© 2010 Wiley Periodicals, Inc.   J Comput Chem 32: 483–496, 2011

Key words: principal component analysis; ligand binding; time series analysis; intraminimum dynamics; vibrational frequencies; collective motions

Introduction

Ligand binding is an important process in controlling protein functions, such as enzymatic reactions, signal transduction, or allosteric effects. Major structural changes have been observed in some proteins upon ligand binging. These structural changes have been usually explained by the induced-fit theorem,[1] which assumes that binding of the ligand brings the protein to an active state; or by the, lately, population-shift model,[2] which asserts that protein, already in the unliganded state, exhibits an ensemble of conformational states, and ligand binding leads to a redistribution of the population of these states. A well-documented classification and discussion of the experimentally determined structural changes in proteins can be found by Goh et al.[3]

Ligand binding does not always lead to remarkable structural changes, but may cause changes in protein dynamics. It has been suggested that ligand binding initiates perturbations propagating to remote sites of the protein.[4] Molecular Dynamics (MD) study of unliganded and liganded Ribonuclease A (RNase A) has shown that substrate binding inhibits the hinge-bending motion, which is present in the unliganded state.[5] X-ray crystallography and MD simulation studies of RNase A have shown that ligand binding leads to reversible subtle domain motions.[6,7] MD simulations of unliganded and two different liganded states of plant nonspecific lipid transfer protein 1 have shown that these three states, though similar in structure, have distinct dif-

Additional Supporting Information may be found in the online version of this article.

*These authors contributed equally

Correspondence to: Burak Alakent; e-mail: burak.alakent@boun.edu.tr

Contract/grant sponsor: TUBITAK; contract/grant numbers: 104M247

Contract/grant sponsor: European 6th Framework Project; contract/grant numbers: EU-FP6-ACC-2004-SSA-2 (contract No. 517991) and Betil Fund

Contract/grant sponsor: DPT; contract/grant number: 2009K120520

ferences with respect to the flexibility of certain residues and global motions.[8] Loop motions and correlation of residues have substantially changed in the ternary complexes of DHFR bound to DHF/NADPH, THF/NADP$^+$, or THF/NADPH, which differ only one or two hydrogen atoms.[9]

Ligand binding has another aspect, namely entropy changes. It has been pointed out that hydrophobic effects, intramolecular vibrations, and conformational entropy are the main sources of entropy changes in the processes involving proteins, specifically ligand binding.[10] Shifting of the vibrational frequencies to higher values or decrease in the number of nearly iso-energetic conformational substrates because of binding results in a decrease in entropy. Binding of a small molecule, while not changing the structure considerably, may alter the overall flexibility of the protein, and so may affect the binding free energy. Comparison of the unliganded and liganded states of protein FKBP-12,[11] acyl-coenzyme A binding protein,[12] and intestinal fatty-acid binding protein[13] by ^{15}N relaxation measurements have shown that liganded states of proteins are less flexible compared with the unliganded states. There is, however, dispute over the flexibility change of the protein because of ligand binding. NMR relaxation experiments have indicated that backbone flexibility of mouse major urinary protein increases upon ligand binding, and this increase is attributed, among other factors, to the increase in the vibrational and conformational entropy terms.[14] Normal mode analysis (NMA) of the unliganded and a water molecule bound structures of pancreatic trypsin inhibitor has indicated a decrease in the vibrational frequencies of the protein, thus vibrational entropy and flexibility of the protein have increased.[15,16] Determination of the change in the vibrational frequencies via neutron scattering experiments of liganded and unliganded states of dihydrofolate reductase (DHFR) at 120 K has shown that vibrational frequencies of the complex shift to lower values.[17] On the contrary, ligand binding to diguanylate cyclase PleD has been shown to shift the normal modes to higher frequencies.[18]

In this study, it is aimed to see how global dynamics and flexibility of proteins change in different ligation states. MD simulations of two different proteins, namely DHFR and triose-phosphate isomerase (TIM) in two different states, unliganded (named as free) and liganded (named as bound), are performed at 300 K. The analyses for DHFR liganded with NAPDH and NADP$^+$ are presented in detail. TIM is used to check the validity of results related with vibrational frequencies, and examine the possibility of generalizing these results to other proteins. The concerted motions of the backbone are analyzed by applying principal components analysis (PCA) on the C$_\alpha$ atomic coordinates from independent MD simulations,[19] whereas dynamics of the principal components (PCs) are determined by using stochastic linear time series models.[20,21] The superiority of stochastic time series models over NMA[22] and quasi-harmonic analysis[23,24] lies in the fact that neither of the latter methods takes into account the anharmonicity of protein dynamics, particularly dominant in the low indexed modes, into consideration. Time series models, on the other hand, separate intraminimum vibrational motions from nonstationarity diffusive motions (random walk motion), and show that ligand binding and small differences in ligands may significantly affect protein dynamics even in ps-time scale.

Materials and Methods

Proteins Used in This Study

The first protein used in this study is DHFR, which catalyzes the reduction of 7,8-dihydrofolate (DHF) to 5,6,7,8-tetrahydrofolate (THF) by using nicotinamide adenine dinucleotide phosphate (NADPH) as the cofactor, and is a target for anticancer and antibacterial drugs. DHFR is bound to NADP$^+$ (oxidized state of NADPH), NADPH, DHF, and THF in binary and ternary complexes on its full kinetic pathway.[25] It has been found that DHFR-NADPH complex lies on the preferred kinetic pathway, and the stability of this complex is considerably higher than DHFR-NADP$^+$ complex.[25] *E. coli* DHFR comprises 159 residues with eight β-strands (strands A to H), four α-helices (helices B, C, E and F), and the loop regions (CD, FG, GH, and M20 loops). The X-ray structures of DHFR have shown that M20 loop (between Ala-9 and Leu-24), whose fluctuations are thought to be important in catalysis,[26] is found in three different ordered conformations: open, closed, and occluded.[27] Figure 1 shows these three conformations and a single molecule of NADPH bound to the closed structure.

In our study, the nicotinamide group has been placed in the active site in both the oxidized and reduced forms. According to X-ray crystallographic studies,[27] this represents the most abundant conformation for the DHFR-NADPH complex. Even though this is a rarely occurring conformation for the DHFR-NADP$^+$ complex, the nicotinamide binding site has a 100% occupancy in the ternary complex with folate or methotrexate.[27] In our calculations, the cofactors have the same conformation as in these ternary complexes, allowing to elucidate their effect on the coupled motions of the protein. As such, we have been able to investigate the effects of the hydride transfer and the formation of a positive charge in the active site on the global motions of the enzyme.

The second protein investigated in this study is TIM, which plays an important role in the fourth step of glycolysis, and is essential for efficient energy production. TIM catalyses the interconversion between dihydroxyacetone phosphate (DHAP) and D-glyceraldehyde 3-phosphate (GAP).[28] TIM comprises 494 residues. Native TIM is active as a dimer. Loop 6 (between Pro-166 and Ala-176 in each subunit) of TIM protects the ligand from solvent exposure during catalysis by closing over the active site. Also, same loop opens and closes in the free (apo) state.

MD Simulation Protocols

MD simulations are performed by using the AMBER 8.0 package[29] together with the ff03 parameters.[30] Simulations are performed in explicit water (TIP3P).[31] A periodic truncated octahedron box of 107 Å dimensions is used for solvation of TIM dimer. In the case of DHFR, a periodic cubic box of 64 Å dimensions is used for solvation. Electrostatic interactions are computed with the particle-mesh Ewald summation technique,[32] where the direct space sum is limited to 9 Å. Energy minimization is performed until the average root mean square gradient has reached 0.01 kcal/mol/Å. Each trajectory is started with velocity assignments according to the Boltzmann distribution at 10 K, and temperature is gradually raised to 300 K. NPT (isothermal-iso-

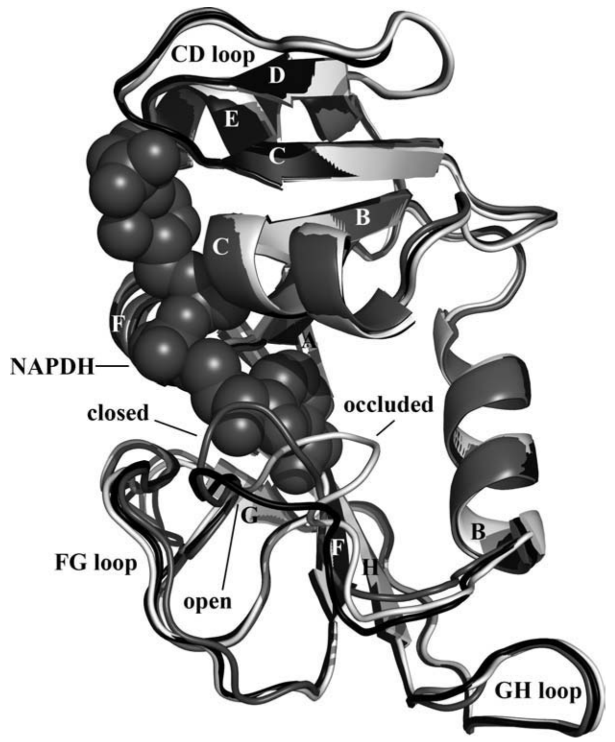

Figure 1. DHFR with M20 loop in open (PDB code: 1RA1, dark color), closed (PDB code: 1RX1, gray color) and occluded (PDB code: 1RX7, light color) conformations, shown in cartoon representation. Nonhydrogen atoms of NADPH are shown in dark grey vdW spheres. The nicotinamide cannot bind in the occluded conformation, but shown for representation purpose. All the molecular graphics images are produced using PYMOL (DeLano Scientific; http://pymol.org).

baric ensemble) simulations are performed at 300 K and 1 bar using the weak coupling algorithm.[33] Constant pressure periodic boundary conditions are used with isotropic position scaling. An integration time step of 2 fs is used by the implementation of SHAKE algorithm for the bonds involving hydrogens.[34]

Three different MD simulations are performed on DHFR, namely the free state (denoted by D), NADP$^+$ bound state (denoted by D$^+$), and NADPH bound state (denoted by D^H). Two different MD simulations are performed on TIM, the free state, and DHAP bound state, denoted by T and T^D, respectively. Protein Data Bank (PDB)[35] is used for obtaining the X-ray crystallography structures, which are used as the starting conformations for the MD simulations. Starting conformation for the DHFR free state is generated by removing the ligand NADPH from the structure with the PDB code 1RA1, because the M20 loop is disordered in the crystal structure for apo DHFR. 1RX1 is used for the ligand-bound DHFRs both in NADP$^+$ and NADPH-bound simulations. The PDB codes for the free and ligand bound states of TIM are 8TIM and 1TPH, respectively. The ligand in 1TPH is an intermediate analog (PGH) having the same structure and orientation with the substrate DHAP, which is used in our simulations. Ryde's force field parameters for NADPH and NADP$^+$ have been used as they are given in the "contributed parameters" link in the AMBER web site (note that the proper reference is not given in this web site). The respective charges for DHAP have been computed using the electrostatic potential obtained from B3LYP/cc-pVTZ calculations in a solvent of dielectric constant of 4. The rest of the parameters have been obtained using the antechamber program of the AMBER package.

Each independent simulation has been performed at least for 20 ns. For example, in the case of apo TIM, simulation length is 60 ns, and a recent detailed analysis of the collective dynamics of TIM is based on this simulation.[28] Each of the long independent simulations is sampled twice (shown by 1 and 2; for example, D^H1 denotes the first sampling of NADPH bound DHFR simulation) from different 3.2 ns long parts, which possibly represent different regions on the potential energy surface, at a sampling interval of 0.8 ps. The reason why short segments are analyzed from long simulations is that the longest period of the vibrational motions (lowest frequency motions) is ~10 ps,[36,37] and a data collection period of ~3 ns is adequate for correctly determining those frequencies. Analysis of longer MD simulations makes it difficult to detect the vibrational frequencies correctly because of the dominance of random walk motions.

PCA

PCA is a statistical method[38] to reduce the dimensions of the trajectory data matrix to capture the essential dynamics of the protein.[19,39] After superimposing all the snapshots of the trajectory onto the initial structure,[40] C_α atomic displacements (**X**) of the protein from the average structure of the simulation are used to calculate the covariance matrix (**C**). Eigenvalue decomposition is applied on this covariance matrix as $\mathbf{C} = \mathbf{P}\Lambda\mathbf{P}^T$, where **P** and Λ are the eigenvectors and the eigenvalues matrices, respectively. It should be remarked that each eigenvector (column vector of **P**) is of unit length, and the eigenvalues (λ) in the diagonal of Λ matrix are in decreasing order. In this case, $\mathbf{p}_1$, the first column vector of **P**, is a linear combination of the C_α atomic displacements with the highest mean square fluctuations (MSF); $\mathbf{p}_{2,}$ the second column vector of **P**, is a linear combination of the variables with the highest variance in the residual covariance matrix, and so on. The projection of the C_α atomic displacements onto ith eigenvector comprises the ith scores vector ($\mathbf{t}_i$), and is determined by $\mathbf{t}_i = \mathbf{X}\mathbf{p}_i$. The variance of $\mathbf{t}_i$ is equal to λ_i.

Collectivity

Collectivity (κ_i) of the ith eigenvector is measured by using the exponential of information entropy[41]:

$$\kappa_i = \frac{1}{N}\exp\left(-\sum_{j=1}^{N}\left(r_{i,j}^2\right)\log\left(r_{i,j}^2\right)\right) \qquad (1)$$

In this equation, N is the number of C_α atoms, and

$$r_{i,j}^2 = p_{i,3j-2}^2 + p_{i,3j-1}^2 + p_{i,3j}^2 \qquad (2)$$

where $p_{i,j}$ is the jth scalar element of the ith eigenvector. It should be noted that normalization factor is not required, because mass weighting is not performed in PCA of C_α atomic trajectories, and the norm of each eigenvector is equal to unity. Thus, $\sum_{j=1}^{N} r_{i,j}^2 = 1$ condition is satisfied.

Linear Stochastic Time Series Models

A discrete time series is a set of observations sampled at fixed time intervals. Statistical time series are results of stochastic processes, and governed by underlying probabilistic mechanisms. A stochastic process may be stationary, that is, having constant probability distribution with respect to time, or nonstationary, such as the level of the process constantly changing. It is possible to model the behavior of nonstationary time series by using autoregressive integrated moving average (ARIMA) processes of order (*p*,*d*,*q*), as[42]

$$\phi(B)(1-B)^d z_t = \theta(B)a_t \tag{3}$$

In this representation, z_t is the mean scaled observation sampled at time *t*; *B* is the backshift operator, such that $Bx_t = x_{t-1}$; $\phi(B)$ is the *p*th order autoregressive (AR) characteristic equation, $\phi(B) = 1 - \phi_1 B - \cdots - \phi_p B^p$; $\theta(B)$ is the *q*th order moving average (MA) polynomial equation, $\theta(B) = 1 - \theta_1 B - \cdots - \theta_q B^q$; *d* is the number of differencing required to make the nonstationary series stationary; and a_t is the independent and identically distributed (usually taken to be Gaussian) random shock term. The variance of random shocks is denoted by σ_a^2. Most of the nonstationary time series can be made stationary with a single differencing ($d = 1$), thus the differencing operator ∇ is used: $w_t = (1-B)z_t = \nabla z_t$. Therefore, an ARIMA(*p*,1,*q*) process can be represented as

$$\left(1 - \phi_1 B - \cdots - \phi_p B^p\right)\nabla z_t = \left(1 - \theta_1 B - \cdots - \theta_q B^q\right)a_t \tag{4}$$

It has been previously shown that it is possible to model the time behavior of the protein collective coordinates by ARIMA(*p*,1,*q*) processes in vacuum[20,21] and in water.[43] Time series analysis is performed by using System Identification Toolbox on MATLAB 7.0 (The MathWorks, Natick, MA).

If the roots of the second-order characteristic equation $(1-\phi_1 B-\phi_2 B^2)$ of a process are complex (underdamped process), frequency of pseudoperiodic motions (f_0) can be calculated by using[42]:

$$f_0 = \frac{1}{2\pi}\cos^{-1}\left(\frac{\phi_1}{2\sqrt{-\phi_2}}\right) \tag{5}$$

Otherwise, the linear process has real roots (overdamped process), and its autocorrelation function decays exponentially with respect to time lags.

Results and Discussion

Analysis of the Average Structure and Residue Mobility of DHFR

The analyzed segments of MD simulations of DHFR in different states are superimposed onto their starting X-ray structures (using the C_α atoms only), and the root mean square displacement (RMSD) of each snapshot to its starting structure is shown in Figures 2a–2c. It is seen that although RMSDs of the free state trajectories are slightly higher than those of the rest of the simulations, all RMSDs are at a constant level of ~1.5 Å. Radius of gyration (r_g) values of the C_α atoms (Figs. 2d–2f) are close to an average value of ~15.5 Å in all simulations, negligibly higher than those of the crystal structures (15.2–15.3 Å). These structural analyses show that MD trajectory structures are stable, and do not indicate any major conformational transitions during the sampling period.

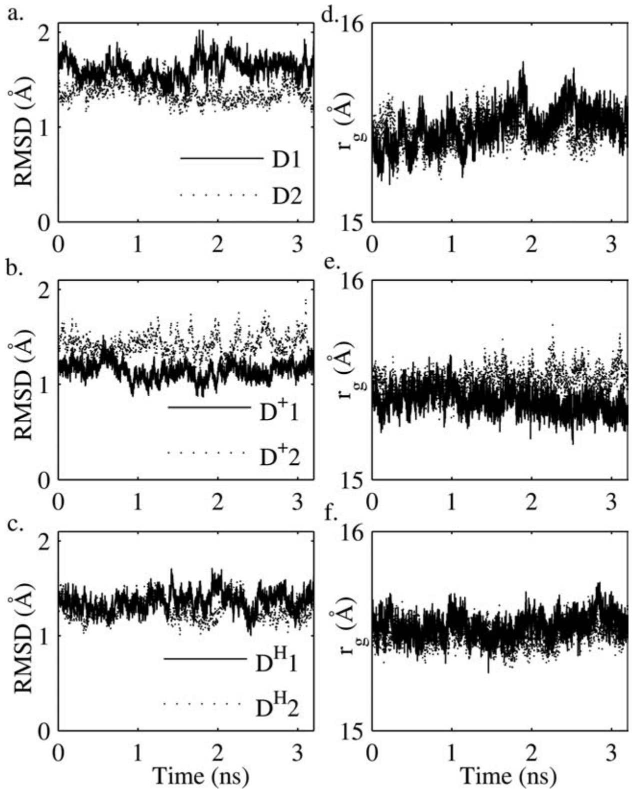

Figure 2. RMSDs of the C_α atoms of each MD simulation to its starting structure: (a) D1 and D2 to 1RA1, (b) D^+1 and D^+2 to 1RX1, and (c) D^H1 and D^H2 to 1RX1. Radius of gyration of the C_α atoms of (d) D1 and D2, (e) D^+1 and D^+2, and (f) D^H1 and D^H2.

RMSD values between the average MD and X-ray structures excluding the M20 loop (see the first line in each cell in Table S1 of the Supporting Information) are about 1 Å for all states, showing that there is no remarkable global conformational difference between the MD simulation conformations and X-ray structures. Average conformations of the MD simulations are shown in Figures S1a–S1c (in the Supporting Information), in which the most remarkable conformational difference (except M20 loop) is seen in GH loop. In D1 and D^+2, GH loop adopts a more extended conformation compared with its crystal structures. On the other hand, existence of this new conformation in both free and liganded states shows that this conformational transition of GH loop is independent of the ligation state.

M20 loop (Ile-14 to Trp-22) is the region showing the most pronounced difference in different simulations. M20 loop in NADP$^+$ bound state is close to the closed conformation, with RMSD values of 1.85 and 1.69 Å for D$^+$1 and D$^+$2, respectively (see the second line in each cell in Table S1). M20 loop in the free and NADPH bound forms of DHFR are close to each other: average conformations of D1 and D^{H}1 are the closest conformations, with an RMSD of 1.06 Å, compared with other combinations in these two forms. Examination of simulation averages show that M20 loop in the free and NADPH bound states adopts an alternative conformation compared with those in the three crystal structures. This alternate conformation is closer (in RMSD value) to the closed conformation, and approximately equidistant to the open and occluded conformations. On the other hand, *K*-means clustering[44] of the pseudodihedral angles of the M20 loop by using MATLAB's statistical toolbox shows that M20 loops in D$^+$1 and D$^+$2 simulations are in the same class with the closed X-ray structure, whereas those in D1, D2, D^{H}1, and D^{H}2 are either in the same class with the open X-ray structure, or form a separate cluster, depending on the number of clusters used.

Residue-averaged MSF are 0.63 (average of 0.71 and 0.55 Å^2), 0.47 (average of 0.45 and 0.48 Å^2), and 0.50 Å^2 (average of 0.51 and 0.48 Å^2) for D1-2, D$^+$1-2, and D^{H}1-2 simulations, respectively. Comparison of MSF of two simulations in each state shows that simulations have reached equilibrium in terms of atomic fluctuations (Figs. S2a–S2c). Examining MSF per residue averaged over two simulations in each state (Fig. 3a) shows that ligand binding reduces the mobility of most residues in the binding region, conforming to the view that ligand binding is accompanied with the loss of translational and rotational entropy terms.[45] Nevertheless, the mobility of the M20 loop is highest in the DHFR-NADPH complex within the time frame of our samples, indicating the complexity of the binding phenomenon.

Analysis of the Essential Dynamics Subspaces Obtained from PCA for DHFR

PCA is applied to all C$_\alpha$ atoms (excluding the ligand) of each of the six MD simulations, separately. Eigenvalues are averaged over the two runs for each state of DHFR (Fig. 3b). The first 10 eigenvalues of D1-2 are the highest, and the first three eigenvalues of D^{H}1-2 are higher than those of D$^+$1-2, proportional to the mobility of the three states. High eigenvalues may be a result of the concerted motion of the protein, or one or more localized regions performing large-scale fluctuations.

Eigenvectors represent the collective motions, and particularly the low indexed eigenvectors, which comprise the essential dynamics subspace,[19] may have functional roles. Similarity of the essential dynamics subspaces obtained from different simulations is examined by the average norm, $\overline{\|u\|}$ with the formulation below:

$$\overline{\|u\|} = \frac{1}{A}\sum_{i=1}^{A}\left(\sum_{j=1}^{A}\left(\mathbf{p}_i\cdot\mathbf{v}_j\right)^2\right)^{\frac{1}{2}} \qquad (6)$$

Here $\mathbf{p}_i$ and $\mathbf{v}_j$ represent the *i*th and *j*th eigenvectors obtained from two different simulations. Eigenvector indices start at one and end at *A*. This formulation is slightly different from the definition of average square norm,[46] with the purpose of making the measure of the overlap of two eigenvector sets quantitatively similar to the dot product of two eigenvectors. In this formulation, $\overline{\|u\|}$ is the average length of the projection of each eigenvector in the first set onto the second eigenvectors subspace. Though the current metric depends on the order of the sets analyzed, application on the data shows that the order of the eigenvector sets on the results is insignificantly small.

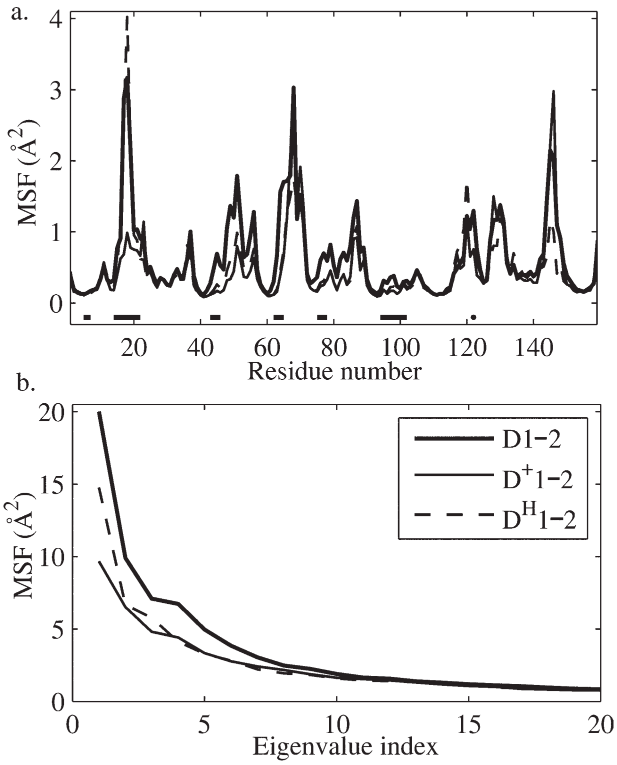

Figure 3. (a) MSF of the C$_\alpha$ atoms obtained by averaging of the simulations of D1-2, D$^+$1-2, and D^{H}1-2. The horizontal lines just below the zero ordinate (the second horizontal line represents the M20 loop) show the residues within a 5 Å distance to the ligand in the X-ray structure. (b) Low indexed eigenvalues averaged over two runs in each of the three states of DHFR.

The overlap values (*A* = 5, 10) within the free and the NADPH bound simulations separately are very close to those between the free and NADPH bound simulations, implying that effect of NADPH on the essential dynamics subspace of DHFR, at first sight, is negligibly small (Table 1). In addition to that overlap values between the essential dynamics subspaces of the D$^+$1-2 and those of the other four simulations are slightly lower than the rest of the overlap values. However, a more detailed analysis of the first five eigenvectors, as shown in the following sections, indicates that these differences are more highly expressed than those predicted from the overlap values.

Analysis of the Domain Motions in DHFR

It has been suggested that DHFR consists of two domains with a rotation axis, which lies between strands A and E. The first do-

Table 1. Average Norms (eq. 6) Between Two Eigenvector Subsets.

	D1	D2	$D^{+}1$	$D^{+}2$	$D^{H}1$
D2	0.67[a]/0.68[b]				
$D^{+}1$	0.59/0.68	0.48/0.63			
$D^{+}2$	0.58/0.69	0.49/0.61	0.56/0.71		
$D^{H}1$	0.63/0.68	0.60/0.68	0.59/0.68	0.57/0.67	
$D^{H}2$	0.67/0.70	0.61/0.66	0.57/0.69	0.48/0.64	0.66/0.73

[a] The first number in each cell represents the average norm when *A* is equal to 5.

[b] The second number in each cell represents the average norm when *A* is equal to 10.

main (named adenosine binding domain, ABD) consists of strands B, C, D, E, helices C, E, F, and loop CD, and the second (larger) domain (named loop domain) comprises strands A, F, G, H, helix B and loops FG, GH, and M20.[27,47] This domain representation is found to be convenient for this study (Fig. 4a), where the domain rotation angles and collectivities are used to measure the extent of rotational motions of the domains along each PC. The relative rotations of domains are determined by the following method. Two snapshots with the minimum and maximum scores attained in the MD simulations along a PC are taken. The two snapshots are superimposed by aligning the large domains. Then, the translation and rotation matrices, required to align the first domain, are computed. The rotation axis is the real eigenvector of the rotation matrix, whereas the rotation angle of the domains is determined by using the trace of rotation matrix. The highest norm of the translation vector is found to be 1.9 Å for PC 1 of D1, whereas the norms of the rest of the translation vectors do not form clusters with respect to ligation states, and their 25 and 75 percent quartiles are 0.55 and 0.90 Å. Thus, translation vectors are not taken into account for the analyses.

Figures 4b–4h show representative displacement vectors (magnitudes oversized by the same scaling factor) along PCs extracted for each ligation state. PC 1, PC 2, and PC 4 of D1 (Figs. 4b–4d) show well-pronounced rotational motions of the two domains about three different and approximately normal rotation axes. The rotation axes determined from the first five PCs of D2 are similar to those of D1 (Fig. S3). A previous study

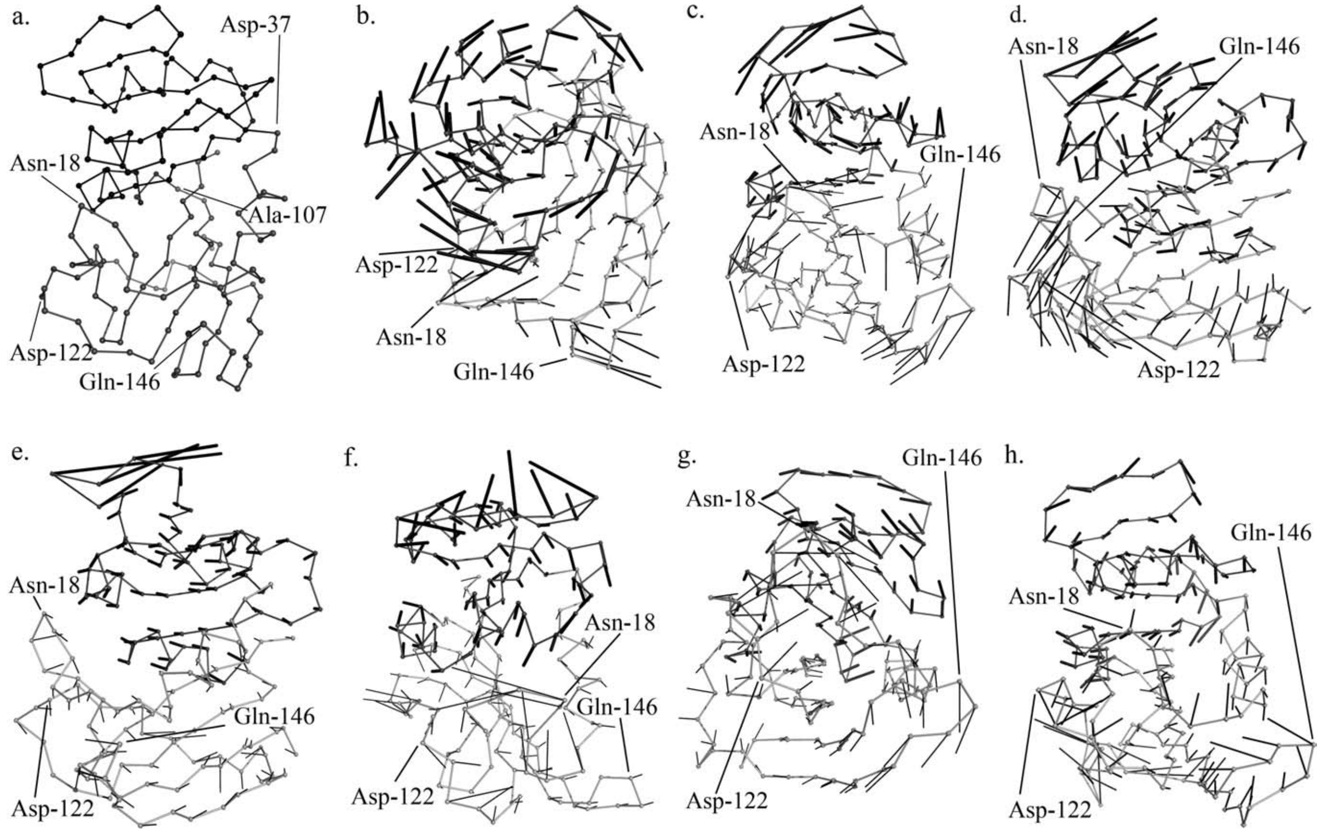

Figure 4. (a) Domain representation of DHFR. C_α displacement vectors on (b) PC 1, (c) PC 2, and (d) PC 4 of D1; (e) PC 1, and (f) PC 2 of $D^{+}1$; (g) PC 1, and (h) PC 2 of $D^{H}1$. In all figures, the lighter colored residues and the thin displacement vectors belong to the large domain, while the darker colored residues and thick displacement vectors belong to ABD. The residues Asn-18, Asp-122, and Gln-146 are shown to give the viewer an idea about the locations of M20, FG and GH loops, respectively. The ABD domain rotation axes are normal to the plane of the page in all figures.

Table 2. Rotation Angles and Collectivities of the First Two PCs and the Averages for the First Five PCs (the last column) in Each MD Simulation.

	PC 1	PC 2	PCs 1 to 5
D1	11.6[a]/0.46[b]	10.2/0.44	9.52/0.45
D2	10.5/0.49	10.8/0.50	7.84/0.45
$D^{+}1$	3.80/0.14	2.63/0.35	4.86/0.38
$D^{+}2$	7.69/0.17	7.33/0.26	5.07/0.35
$D^{H}1$	10.4/0.26	3.92/0.46	6.00/0.36
$D^{H}2$	9.75/0.33	8.41/0.37	6.50/0.46

[a]The first number in each cell is rotation angle in degrees.

[b]The second number in each cell is collectivity.

of a tertiary complex of DHFR has shown that the two domains rotate about two perpendicular axes.[47] Considering that MD simulation period in that study was 208 ps, we believe that the current analysis gives a more complete picture of the domain motions in DHFR. The global domain rotations are replaced by a local motion of CD loop along the PCs 1 and 2 of the DHFR-$NADP^{+}$ complex (Figs. 4e–4f). Domain rotations are captured, but the global contribution of residues to the domain rotation is reduced in the presence of NADPH, mostly because of the high contribution of the M20 loop and the low contribution of ABD to PC 1 (Fig. 4g). Domain rotation is almost completely lost along PC 2 (Fig. 4h).

It is also important to note that dot product of PC 4 of D1 (Fig. 4d) with PC 1 of $D^{+}1$ (Fig. 4e) is 0.49, and their rotation axes make an angle of 22.3°. One would, naively, expect that rotational motions on these PCs are similar, but one can hardly see a domain rotation in PC 1 of $D^{+}1$. On the contrary, dot product of PC 2 of D1 (Fig. 4c) with PC 1 of $D^{H}1$ (Fig. 4g) is a rather low value at 0.23, but these two eigenvectors represent a similar domain motion, with an angle of 23.4° between their rotation axes.

The visual examination presented above is verified and summarized by the rotation angles and collectivities of the first five PCs shown in Table 2. Low indexed PCs of the free state have high rotation angles and collectivities, indicating that a large number of residues contributes to the rotational motion. Domain rotations are considerably reduced in the DHFR-$NADP^{+}$ complex, with both the rotation angles and collectivities damped, whereas the domain rotations in the DHFR-NADPH complex lie between those of the free and $NADP^{+}$ bound states.

To check the validity of the analysis above, backbone atom coordinates are generated from C_{α} traces by using MaxSprout[48] (http://www.ebi.ac.uk/Tools/maxsprout) and domain motions described by the first five PCs are reexamined via using DynDom[49] (http://fizz.cmp.uea.ac.uk/dyndom). It is seen that our domain analysis agrees very well with that of DynDom. Among the first five PCs of D1 and D2 simulations, DynDom finds six dynamic domains, which closely resemble ABD and large domains, with both twisting and closure type of motions of rotation angles between 10 and 13° (Table S2). There is only one PC with significant domain motion among the first five PCs in the $NADP^{+}$ bound state simulations (Table S3), confirming the significant distortion of domain motions upon binding of $NADP^{+}$. In NADPH bound state, there are four PCs with dynamic domains (Table S4). Two of those correspond to ABD and the large domain, but they represent only twisting type of motion with smaller rotation angles of 9–10°. In the other two PCs, dynamic domains are confined to local regions, suggesting the dampening of domain motions in the NADPH bound state.

Difference of domain motions in DHFR-NADPH and DHFR-$NADP^{+}$ complexes can be partially explained by the difference in the electrostatic interactions between the cofactor and DHFR in these two complexes. The nicotinamide ring of the cofactor is surrounded by four oxygen atoms: backbone oxygens of Ile-14 and Ile-94, and side chain oxygens of Thr-46 and Tyr-100. The positively charged ring of $NADP^{+}$ interacts favorably with the partial negative charges on these oxygen atoms. On the other hand, the uncharged ring of NADPH is not expected to interact significantly with these oxygens. Average distances, based on atomistic snapshots, between each oxygen and the closest hydrogen of the nicotinamide ring have been calculated for both $NADP^{+}$ and NADPH. The average distance between the Tyr-46 oxygen and the nicotinamide ring hydrogen is 2.63 and 2.49 Å for NADPH and $NADP^{+}$, respectively. Similarly, the distance between the Ile-14 oxygen and the ring hydrogen is 2.44 and 2.38 Å for NADPH and $NADP^{+}$, respectively. Also, the distance between the Ile-94 oxygen and the ring hydrogen is shorter with $NADP^{+}$ (2.64 Å) than with NADPH (2.93 Å). Only with Tyr-100, the interaction distance is longer with $NADP^{+}$ (2.85 Å) than with NADPH (2.69 Å). The shortening of most of the interactions can be taken as an evidence for the strengthening of the electrostatic interactions between the cofactor and the protein with $NADP^{+}$ relative to NAPDH.

Analysis of the DHFR Motions by Time Series Analysis

Fitting Time Series Model to a Scores Trajectory

The $\mathbf{t}_3$ scores trajectory of D1 is given as a representative example to demonstrate how a time series model is fitted to a MD trajectory. The normalized $\mathbf{t}_3$ scores in Fig. 5a (denoted by z_t) show that the time series is nonstationary, and the distribution of the scores is not Gaussian (not shown). Autocorrelation function of the scores (Fig. 5b) does not die out and levels off ~0.7, indicating the nonstationarity of the process. Application of a single differencing operator (∇) on the scores results in a time series (w_t) with a constant mean (Fig. 5c) and Gaussian distribution (not shown). This shows that though the average position of the protein along the mode (PC) varies in time, displacements between conformations sampled at 0.8 ps follow a stationary trajectory. The mean of the series w_t (−0.002 Å) is much smaller than the 95% confidence limit (0.024 Å), showing that displacements in equal magnitudes are likely to occur in either direction. Autocorrelations of the displacements (Fig. 5d) at the first three nonzero lags are higher than the "large lag" 95% confidence limits, and autocorrelation at the first time lag is smaller (in absolute) value than the other two, which seem to show a decaying behavior. These indicate that the time series model should contain both AR and MA parameters.[42] Model order selection is described in the Supporting Information and Figure S4, and

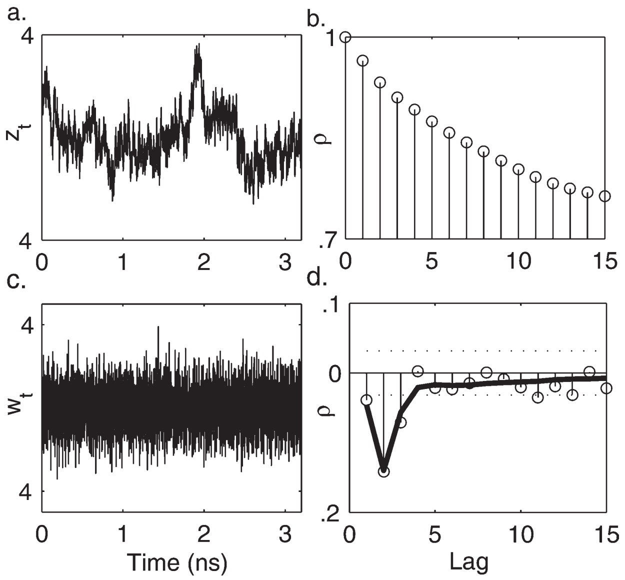

Figure 5. (a) The t_3-scores trajectory (z_t), normalized with respect to the standard deviation of the time series. (b) Sample autocorrelation function (ρ) of t_3-scores with respect to sampling lags of 0.8 ps. (c) Difference of successive t_3-scores, normalized with respect to standard deviation. (d) Sample autocorrelation function (shown with dots) of the w_t series. Dashed lines represent the large lag 95% confidence limits for zero autocorrelation. Thick solid line shows the theoretical autocorrelation function, obtained from the ARMA(3,2) model.

ARMA(3,2) is found to be a satisfactory model. The theoretical autocorrelation function obtained from the ARMA(3,2) model is shown with solid lines in Figure 5d, and conforms well with the simulation results.

It should be recalled that ARMA model is derived for w_t series, which has been obtained by differencing z_t series. Thus, z_t series is modeled by ARIMA(3,1,2) model, shown in the operator representation as follows:

$$(1-0.902B)\left(1-0.294B+0.104B^2\right)\nabla z_t = (1-0.971B)(1-0.311B)a_t \quad (7)$$

where z_t represents the $\mathbf{t}_3$ scores. Our previous studies[20] showed that the complex roots of $(1-0.294B+0.104B^2)$ term represent the vibrational motions, while one of the MA terms $(1-0.311B)$ is to account for the autocorrelation of random force term. The $(1-B)$ term on the left hand side represents the nonstationary interminimum motions. The $(1-0.971B)$ and $(1-0.902B)$ terms on both sides, though close to each other, do not cancel out. In our previous studies,[20,21,43] ARIMA(2,1,1) or ARIMA(2,1,2) processes have been sufficient for modeling the dynamic behavior along the principal modes. However, in this study, most of the scores trajectories for DHFR and TIM, which are comparatively larger proteins, can be modeled by ARIMA(3,1,1) and ARIMA(3,1,2) processes. Therefore, a higher number of AR terms may be necessary to approximate the nonlinear behavior of larger proteins by linear models.

Comparison of Conformational Subspaces of DHFR in Different States

All six MD simulation trajectories are projected on the same subspaces formed by the first five principal modes obtained from each MD trajectory, and time series models are constructed for the projections of all simulations on the same PCs. It is important to realize that the two repetitions of the MD simulations in each state will enable us to make this comparison without any bias stemming from PCA.[50] For instance, D2, D^H1, and D^H2 trajectories are all to be projected on the same PC subspace obtained from D1 trajectory, and a difference observed in the time series model parameters of D2 from those of D^H1 and D^H2 can be attributed to an actual difference in the protein dynamics between the free and NADPH bound states.

Figure 6a shows the standard deviations (equal to root mean square fluctuations) of the projections of D1, D2, D^H1, and D^H2 trajectory on each of the first five PCs from D1 simulation. Root mean square fluctuations (RMSF) of the projection of MD_1 on PC 1 are significantly higher than the rest of the RMSF of the projections because PCs are obtained from D1 simulation. On the other hand, one cannot distinguish D2, D^H1, and D^H2 with

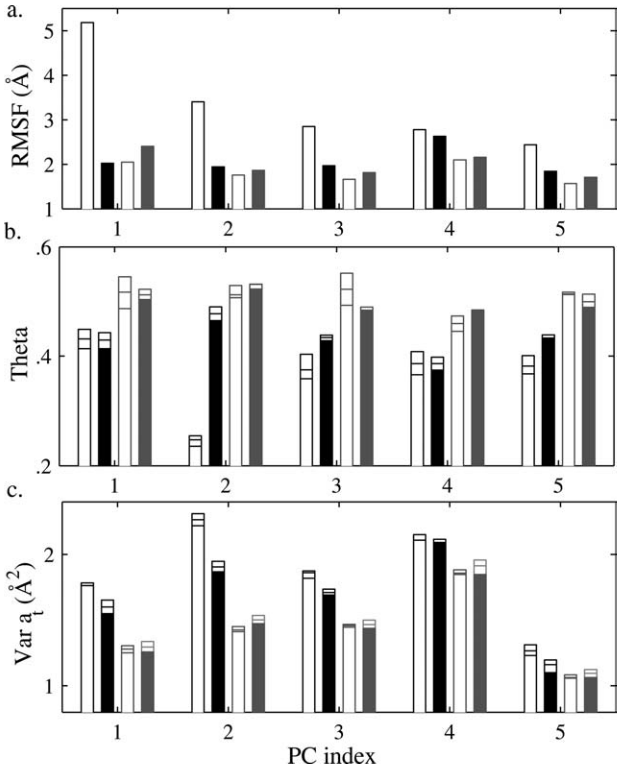

Figure 6. Projections of MD simulations in free and NADPH bound states on the first five PCs of D1. (a) RMSF along the modes, (b) θ parameter, and (c) σ_a^2 values in the IMA(1,1) models. Empty black, filled black, empty gray and filled gray boxes represent D1, D2, D^H1, and D^H2 simulations, respectively. Horizontal lines represent lower percentile (25%), mean and upper percentile (75%) of the time series model parameters in the five different models obtained by sliding the sampling window by 0.8 ps.

respect to RMSF along the PCs. Time series models are used to make a more detailed and reliable comparison of the dynamics in different ligation states.

In the following analysis, two sets of time series models are constructed. The first one is a low order IMA(1,1) time series model, and the second one is a higher order ARIMA(3,1,2) model. Sampling interval is increased from 0.8 to 4 ps (every five samples are taken) to fit IMA(1,1) model to the MD trajectories in a statistically satisfactory manner. Although less frequent sampling would cause information loss on the details of vibrational and relaxation processes (mostly intraminimum motions) at short time intervals, the purpose of constructing IMA(1,1) models is to simplify the analysis of dynamics.

The IMA(1,1) processes are also known as exponentially weighted moving average (EWMA)[51] processes, and shown as:

$$\nabla z_t = (1 - \theta B) a_t \qquad (8)$$

If θ is close to zero, then the process approaches to a random walk. As θ is closer to unity, the process approaches to a stationary process. Steps taken between successive process levels are proportional to σ_a^2 values. Therefore, a small value of θ and/or a large value of σ_a^2 characterizes a mode with high mobility.

Five IMA(1,1) models for each trajectory are obtained by sliding the sampling window by 0.8 ps, and the averages (of the estimates obtained for each sampling window) of θ parameters (Fig. 6b) and σ_a^2 values (Fig. 6c) for each of the first five individual modes are shown in bar graph representation with lower and upper percentile values. Considering the low variation of the parameter estimates, one can safely say that θ values are smaller and σ_a^2 values are larger in the free state compared with NADPH bound state.

The above analysis is repeated by projecting all simulations to each of the PCs 1–5 subspaces obtained from D1, D2, that is, PC subspace of the unliganded protein simulations. The projected trajectories are modeled by IMA(1,1) processes, and time series model parameters θ and σ_a^2 are averaged over the (first five) PCs to make PC subspace comparisons easier (Table 3). For instance, the first column in Table 3 (under the columns with the title "PCs of D1") corresponds to the results in Figures 6b and 6c averaged over the first five PCs. The difference between the time series parameters of D1 and D2 simulations, that is, θ averages are found to be 0.36 in D1 and 0.43 in D2, can be attributed to PC 2 in D1, showing a significant high mobility compared with the rest of the PCs (Figs. 6b and 6c), and to the conformational heterogeneity[52] of the highly flexible free state. Similar to the subspace of D1, in the subspace of D2 (under the columns with the title "PCs of D2"), θ and σ_a^2 parameters take significantly different values for liganded and unliganded states: the average θ parameter is 0.38 for both D1 and D2, whereas the average of θ is found to be ~0.50 for the liganded simulations. The two states can also be distinguished based on their σ_a^2 values. In other words, short time dynamics (4 ps) of the liganded states is perturbed in such a way to decrease the mobility of DHFR along the modes representing domain rotations. Although the difference between the parameter values of the NADP^+ and NADPH bound states is less pronounced, for example, the σ_a^2 values in the NADP^+ bound state are smaller than those in the NADPH bound state, but with overlapping confidence limits, repeating the analysis in the PC subspaces of NADPH bound states shows that difference between the two states persists (Table S5). These results suggest that there may be an actual difference in the dynamics of the NADP^+ and NADPH bound states, and the the ps-scale dynamics show that domain rotations of DHFR will be the least pronounced in the NADP^+ bound state.

Table 3. Averages and Confidence Limits of the Model Parameters θ and σ_a^2 of the IMA(1,1) Models Fitted to the MD Trajectories Projected on the First Five PCs of the Free State.

	PCs of D1[a]		PCs of D2	
	θ	σ_a^2 (Å^2)	θ	σ_a^2 (Å^2)
D1[b]	0.36 (0.03)[c]	1.85 (0.14)	0.38 (0.03)	1.65 (0.14)
D2	0.43 (0.01)	1.70 (0.13)	0.38 (0.02)	1.68 (0.16)
D^+1	0.50 (0.03)	1.32 (0.09)	0.53 (0.02)	1.19 (0.11)
D^+2	0.50 (0.02)	1.37 (0.14)	0.53 (0.02)	1.17 (0.12)
D^H1	0.51 (0.01)	1.42 (0.11)	0.51 (0.02)	1.29 (0.13)
D^H2	0.50 (0.01)	1.45 (0.11)	0.49 (0.02)	1.34 (0.13)

[a] The two columns correspond to different subspaces, comprising the first five PCs of D1, and D2 simulations, respectively.

[b] Each row corresponds to a different simulation projected on the PC subpaces. Comparisons between projected simulations should be done between the rows as column by column basis.

[c] The first number and the number in parenthesis are the parameter averages and two standard deviation limits, respectively, obtained from the first five PCs.

Dynamics of the modes in a shorter time scale are examined by using ARIMA (3,1,2) models on the same trajectories. Large number of parameters in ARIMA (3,1,2) model renders its comparison difficult, so ARIMA model is converted into an infinite impulse response (IIR) model, via polynomial division of the MA equation by the AR equation. As a representative example, the first seven terms of IIR model, obtained from the ARIMA model of the $\text{D}^\text{H}2$ trajectory projected on PC 1 of D1, are shown in the following equation:

$$\nabla z_t = (1 - 0.143B - 0.186B^2 - 0.097B^3 - 0.060B^4 \\ -0.048B^5 - 0.041B^6 - \cdots) a_t \qquad (9)$$

with σ_a^2 value of 0.448 Å^2. Coefficients are negative at all time lags (except at lag zero) and drop down (in absolute value) to 0.01 in ~14 time lags (~10 ps). Negative IIR coefficients at lags greater than zero show that the effect of an impulse input (random shock) on the displacements at lags greater than zero is in the opposite direction to its initial (zero lag) effect. Therefore, modes, which have smaller (in absolute value) IIR coefficients, should have higher mobility.

In the free and NADPH bound states, averages and lower and upper percentiles of the IIR coefficients (θ_i) with respect to the first five PCs are shown for the first five time lags in Figure 7a. It is seen that the average IIR coefficients are smaller in absolute value in the free state compared with NADPH bound

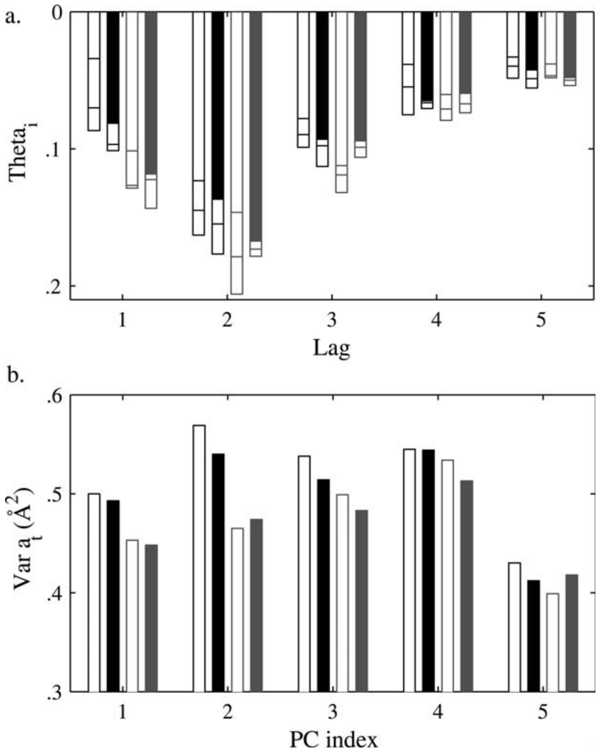

Figure 7. (a) IIR coefficients (θ_i) averaged over the first five PCs of D1 with respect to the first five time lags (0.8 ps), and (b) σ_a^2 values with respect to the first five PCs. Shading and bar representation are identical to those in Figure 6.

state for the first two time lags (0.8 ps and 1.6 ps). Furthermore, variance of random shocks on the first five PCs are higher for the free state compared with NADPH bound state (Fig. 7b). It has been previously shown[20] that difference in the variances of random shocks is due to the difference in vibrational frequencies, and the analysis in the following section will confirm this. Therefore, time series models of the free and liganded states show difference even at 0.8 ps sampling interval, indicating that shape of the local minima, which determine the intraminimum motions, on the energy landscape is changed due to ligand binding.

Determination of the Vibrational Frequencies of DHFR

The change in IIR coefficients and σ_a^2 values upon ligand binding implies that vibrational frequencies are lower in the free state. Equation 5 is applied on the characteristic equations of the time series models to determine vibrational frequencies. Frequencies of the time series models on the PC 1 of D1 are found to be 1.14, 3.88, and 7.18 cm^{-1} for D2, D^{H}1, and D^{H}2 simulations, respectively, whereas the characteristic equation of D1 is found to be overdamped. Projecting all the simulations to PC 1 of D2 gives time series models with frequencies of 1.14, 1.94, and 5.66 cm^{-1} for D1, D2, and D^{H}1 simulations, respectively, and the model for D^{H}2 is found to be overdamped. Unfortunately, higher indexed modes among the first five are found to be overdamped, and vibrational frequencies cannot be obtained. Difficulty of determining vibrational frequencies on the lowest indexed PCs has been reported also in other studies.[53] This is mainly due to water's enhancing of diffusive motions in proteins.[43,54] Therefore, the low frequency vibrational spectra of DHFR in free and NADPH bound states are determined in their own principal mode subspaces. The first 60 scores trajectories are modeled to obtain a sufficient number of vibrational modes for reliable comparison.

Number of underdamped modes among the first 60 modes (24 and 13 for D1 and D2, respectively, whereas 10 and 22 for D^{H}1 and D^{H}2, respectively) does not give a clear indication whether ligand binding has an effect on the dampening of protein's collective vibrational motions in DHFR. The low vibrational frequencies obtained from both simulations of the same states are used to plot histograms with bin widths of 1.5 cm^{-1} (Figs. 8a–8b). There is a remarkable difference between the free and NADPH bound forms of DHFR. There are three modes with vibrational frequencies lower than ~3 cm^{-1} in the free state, whereas there is no mode with frequency lower than this value in the NADPH bound state. There are 18 and 7 PCs with vibrational frequencies lower than 6.25 cm^{-1} in the free and NADPH bound states, respectively. The frequency distribution of the free state has a peak around ~5 cm^{-1}, whereas NADPH bound state has a flatter distribution. It should be reminded that adding PCs with indices higher than 60 would not change the lowest frequency spectrum, which is focused on in the current analysis.

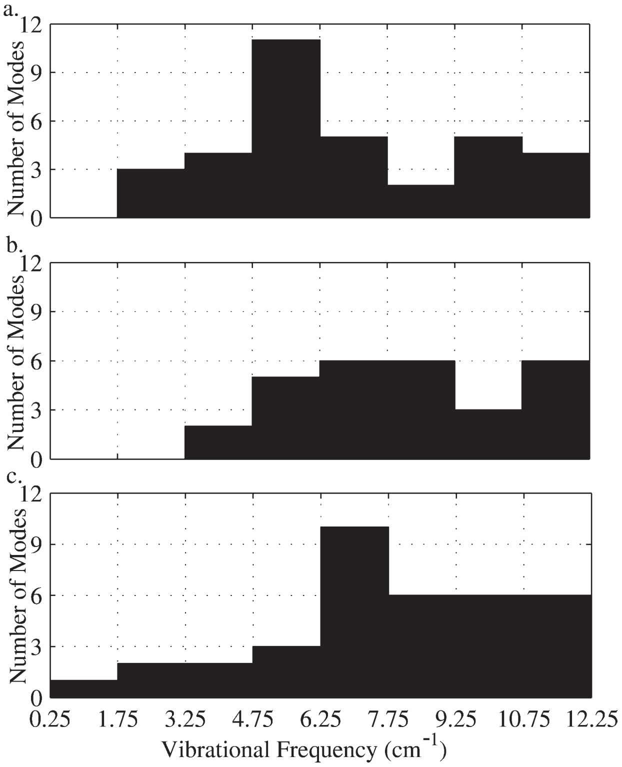

Figure 8. Histograms of the vibrational frequencies (cm^{-1}) of (a) DHFR in the free state, (b) DHFR in the DHFR-NADPH complex, and (c) DHFR-NADPH complex with additional nodes from NADPH. Histograms are plotted with bins of 1.5 cm^{-1}.

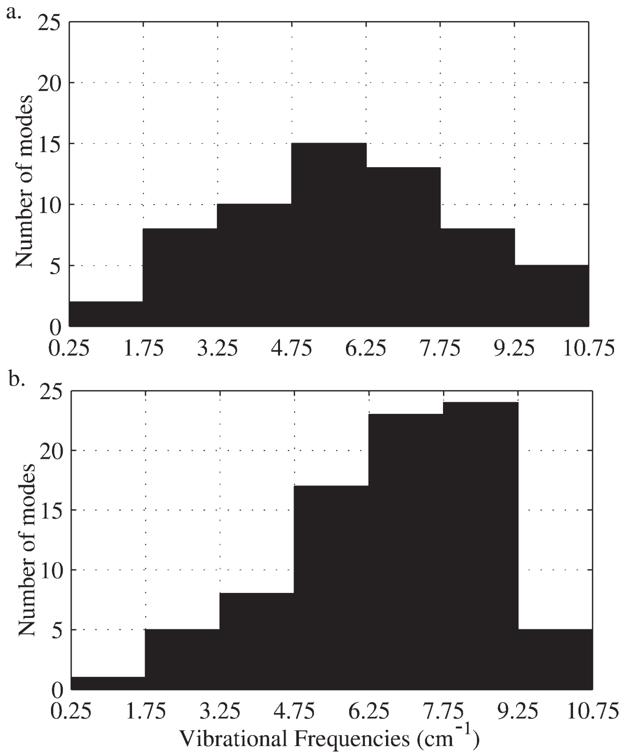

Figure 9. Histograms of the vibrational frequencies (with bins of 1.5 cm^{-1}) of the (a) free TIM (T1 and T2 simulations), and (b) TIM in the TIM-DHAP complex (T^{D}1 and T^{D}2 simulations).

Vibrational frequencies of the DHFR-NADPH complex are obtained by applying PCA on the C$_\alpha$ atoms of DHFR combined with additional seven nodes from the heavy atoms of NADPH, to keep the number of atoms per node similar in protein and ligand. Linear time series models and vibrational frequences are obtained for the 60 PCs of the complex. Histogram of the vibrational frequencies (Fig. 8c) shows that NADPH binding lowers the vibrational frequencies of the complex. There are eight PCs with frequencies lower than 6.25 cm^{-1} (equal to the number of vibrational frequencies obtained for DHFR in the DHFR-NADPH complex in this frequency range), but these frequencies are shifted to lower values, even lower than 2 cm^{-1}. The lowest vibrational frequencies of a protein are proportional to its mass,[55] thus this result shows, once again, the reliability of time series analysis tools on protein dynamics.

Effect of Ligand Binding on the Vibrational Frequencies of TIM

The same procedure followed in DHFR analysis is applied on the data obtained from the simulations of TIM. The simulations T1 and T2 are sampled from the open and closed states of active site loop 6 (both in the absence of ligand), whereas T^{D}1 and T^{D}2 are sampled from the DHAP bound state. Free state has a higher mobility than ligand bound state, with residue averaged MSF values of 0.68 and 0.62 Å^2, respectively. Though the detailed results are not given in this study, it should be mentioned that the rotational motion of two monomers and the coupling of loop 6 with the global motions can be captured by the low indexed PCs for both the free and liganded states.

Time series models of the first 60 PCs are derived as stated in the previous section. ARIMA(3,1,1), ARIMA(3,1,2), and ARMA(3,1) processes are the most commonly encountered models, similar to DHFR. Unlike to what is observed for DHFR, ligand binding increases the number of underdamped modes (37 and 25 for the free states, whereas 43 and 44 for the liganded states) in TIM, significantly. Probability distribution functions of the vibrational frequencies obtained for the open and closed forms (T1 and T2) are similar (Fig. S5), showing that a loop opening/closing, which is essentially sampling of a different region of the energy landscape, does not have a remarkable effect on the collective vibrational frequencies. Comparison of the histograms of the vibrational frequencies obtained from the free and liganded states (Figs. 9a and 9b) shows that the free state, similar to what is observed for DHFR, has a higher number of modes in the lowest frequencies compared with the liganded state. There are 20 and 14 PCs with vibrational frequencies smaller than ~5 cm^{-1} in the free and liganded states, respectively. On the other hand, unlike the flat distribution of DHFR in the DHFR-NADPD complex, probability density function of the vibrational frequencies of the liganded TIM is shifted to higher values. It should also be noted that vibrational frequencies of TIM are lower compared with DHFR, because TIM has a higher molecular mass.

Conclusions

Previous studies in the literature have shown that C$_\alpha$ atoms are responsible from the essential internal motions[19], and analyses based solely on C$_\alpha$ atoms have been successful in predicting thermal fluctuations,[56] vibrational frequency distributions, and collective motion directions in proteins.[57] Therefore, in this study, we focus on C$_\alpha$ atomic trajectories, which are representative of the functionally important collective backbone dynamics, and examine the global motions of DHFR upon ligand binding in detail.

PCA is applied on the MD trajectories, and the average norms of the essential dynamics subspaces of different ligation states show small differences. Nevertheless, the similarity of the essential dynamics subspaces should not make one think that ligand binding has marginal effects on concerted protein motions. In addition to that one should also be cautious with the dot products of the eigenvectors in different MD simulations: relatively high dot products may be obtained between eigenvectors, which do not always represent similar motions. On the other hand, low dot products may be obtained for eigenvectors, which sometimes represent identical domain rotations. Therefore, a detailed quantitative and visual examination of domain motions is required to compare the concerted motions in different simulations.

It has been found previously that correlations of the residues in DHFR vary in different tertiary complexes, which differ only by a hydride ion.[9] In this study, we show that this change is observed in the NADP$^+$/NADPH bound states, even in the absence of the substrate. Upon binding to NADP$^+$, domain rotation of DHFR is remarkably damped, and loop domain is divided into subdomains. Local regions, such as CD, GH, and

M20 loops, dominate the motions along the modes, and this decreases the collectivity of the lowest indexed PCs. Rotations are less pronounced in the DHFR-NADPH complex compared with free state, but more defined compared to NADP$^+$ bound state. It has been asserted that directionality of vibrations affects the substrate specificity,[58] and ligand induced protein conformational flexibility may be important in the co-operativity in the ligand binding.[59] The rotational domain motions we observe are possibly important in substrate/cofactor binding/releasing steps, because they have direct influence on the area and shape of the cofactor and substrate binding sites.

Linear stochastic time series models are used to characterize the dynamics on principal modes in different ligation states. Vibrational and diffusional dynamics of proteins have been successfully described by various models in the literature, such as jumping among minima (JAM)[39] model, diffusion between multiple harmonic wells,[46] moving normal mode coordinates,[60] or probabilistic diffusion-vibration Langevin descriptions.[53] It should be noted that most of these approaches assume a priori model, which is derived either from another model, such as Langevin equation, or a plausible mathematical function, and fit the model to mean square displacements, position, or velocity autocorrelation functions obtained from MD trajectories. Stochastic time series models, on the other hand, are black box models[61] and constructed via minimizing residuals with respect to model parameters. Autocorrelation function of the residuals of a statistically acceptable time series model should be within confidence limits at lags greater than zero. Determining the model order by this method is statistically more reliable than fitting a theoretical autocorrelation function to a sample autocorrelation function.[42] Another advantage of time series models lies in its ability of handling the diffusional characteristics of the protein. Application of the difference operator transforms the collective coordinates (scores) to displacements during the sampling interval, and unlike the collective protein fluctuations, which do not converge during MD simulations,[62] the resulting displacements trajectory is a stationary series, on which statistical analyses can be used. It is important to note that autocorrelation function of the differenced coordinates is similar to velocity autocorrelation function (compare Fig. 4 in ref. 53 and Fig. 5d in this study). On the other hand, as shown in this study, adjusting the sampling interval for differencing brings considerable flexibility in modeling protein dynamics in different time scales.

A detailed time series analysis of the first five modes, describing the domain rotations, has been performed by projecting all MD simulations on the same principal mode subspaces. Two different time series models, ARIMA(3,1,2) and IMA(1,1), at different sampling intervals are used to characterize the dynamic behavior of different ligation states along the same principal modes at picosecond and subpicosecond scales. The slowest frequencies in proteins, such as bovine pancreatic trypsin inhibitor, lysozyme, ribonuclease I, or crambin, are known to be ~2–3 cm^{-1} (corresponding to a period of 10–15 ps),[36,37,55] and most of our discussion on the vibrational spectrum of different ligation states (Figs. 8 and 9) focuses on frequencies lower than ~6 cm^{-1} (~5.5 ps). To construct ARIMA(3,1,2) models, which are capable of characterizing both the diffusional and the vibrational dynamics, data are sampled at a sampling interval of 0.8 ps. The corresponding Nyquist frequency[61] is ~21 cm^{-1}, which is much larger than our bandwidth of interest (the lowest frequencies, 0–6 cm^{-1}). It should also be remarked that faster sampling would cause AR parameters to be pushed to unity[61], and make time series models less reliable. The shortcoming of ARIMA(3,1,2) models, on the other hand, is the large number of parameters, which make the comparison between different simulations difficult. A lower order model, such as IMA(1,1) with a slower sampling rate, is more convenient for this task. Although pseudoperiodic motion cannot be captured by IMA(1,1) models (an AR polynomial of at least order two is required for that), IMA(1,1) processes can be characterized with only two parameters, making the comparisons easier. IMA(1,1) models are used extensively in statistical process control,[51] and describe the current output as an exponentially weighted average of its previous observations and the current random error. In terms of protein motions, IMA(1,1) model assumes that protein moves randomly between different "levels" along the modes, and the MA parameter and variance of the random error term give a measure of the spread of these levels. In other words, mobility along the modes can be estimated by examining the model parameters. After trial and error, the shortest sampling interval adequate for the construction of IMA(1,1) models with satisfactory statistical properties of residuals is found to be 4 ps. Sampling interval is taken to be as small as it can be, so to focus on the intraminimum dynamics of the protein fluctuations. A larger sampling interval could also be used to construct IMA(1,1) processes, but then diffusional interminimum dynamics would dominate the protein fluctuations. Theoretically, the fastest frequency that can be determined at the sampling period of 4 ps is ~4 cm^{-1}, and most of the vibrational frequencies on the modes exceed this value, justifying the use of the model and the sampling interval.

Although realizations of the stochastic processes in different states are indistinguishable, that is, RMSF along the modes are not different in different states, time series analysis shows remarkable differences in the underlying processes. Time series models show that backbone flexibility of DHFR in three states are ranked, in descending order, as free state, NADPH bound state, and NADP$^+$ state, in accordance with the MD simulation results. It should be emphasized that time series models are based on ps-scale sampling, thus atomic displacements in this time range are modeled. It is interesting that mobility of the modes in different ligation states is manifested at such a short time interval. A recent study has shown the μs-ms timescale backbone dynamics of DHFR can be significantly changed by minor chemical or structural differences in ligands.[63] Our study supports this finding, showing that the long time behavior (interminimum motions) information of a protein is buried in its short time (intraminimum motions) dynamics; ligand binding and minor differences in ligands may perturb the intraminimum motions, and time series models can successfully capture these differences in dynamics.

To gain more insight about the intraminimum dynamics, vibrational frequencies are examined using time series models. As the first five principal modes are found to be overdamped, we find it necessary to include the first 60 PCs in our analysis to obtain a sufficient number of vibrational frequencies for a reliable comparison between ligation states. The main difference of our method in determining vibrational frequencies from the conventional Fourier transform of velocities[53] lies in our focus on the large scale collective motions. We use time series analysis on the principal modes representative of large-scale concerted motions to determine the underlying dynamics in short time scales. Fourier transform, on the other hand, extracts the frequency spectrum of all atomic velocities without taking the collective motions into regard. It should also be noted that the current method enables us to focus on the large-scale collective motions of the sole protein in the protein + ligand complex. We think that our findings shed light on the dispute in the literature on how vibrational frequencies are affected upon ligand binding.[11–18] Examination of DHFR with the ligand's atoms taken into consideration shows that lowest vibrational frequencies detected for the DHFR-NADPH binary complex are lower than the unliganded DHFR. This confirms a previous experimental study,[17] in which vibrational frequencies of DHFR are found to shift to lower values for the protein + ligand complex. However, further analysis shows that this difference is due to the additional mass of the ligand, because the lowest vibrational frequencies obtained only for DHFR in the DHFR-NADPH complex are higher than those of apo-DHFR. Similar to that seen in DHFR, unliganded TIM has lower vibrational frequencies compared with its liganded state. On the other hand, DHFR in the liganded state has a flat low frequency distribution, whereas the low frequencies of TIM in the liganded state are shifted to higher values. The difference seen in the frequency distributions of DHFR and TIM may reflect an actual difference in the underlying dynamics of the proteins, or may be due to the small sampling size of the modes, reduced as a consequence of water's dampening of vibrational motions. Application of the current method to different proteins + ligand complexes can answer this question.

The main difficulties in analyzing protein fluctuations arise from the nonlinearity of protein dynamics, superposition of intraminimum and interminimum motions, which take place in remarkably different time scales, and water's effect on these motions. We have shown that linear time series models can successfully characterize the differences in protein dynamics with respect to different ligation states, which do not have remarkable structural differences. As of future study, we would like to improve the efficiency of this method. We are currently working on the optimization of sampling interval and total simulation period to obtain parsimonious and robust models, and application of different filters to the MD trajectories to select desired frequency ranges.

Acknowledgments

The authors thank Dr. Bulent Balta for providing simulation data and valuable discussions.

References

1. Koshland, D. E. Proc Natl Acad Sci USA 1958, 44, 98.
2. Kumar, S. M.; Buyong, M.; Tsai, J. C.; Sinha, N.; Nussinov, R. Protein Sci 2000, 9, 10.
3. Goh, C. -S.; Milburn, D.; Gerstein, M. Curr Opin Struct Biol 2004, 14, 104.
4. Freire, E. Proc Natl Acad Sci USA 1999, 96, 10118.
5. Nadig, G.; Vishveshwara, S. Biopolymers 1997, 42, 505.
6. Vitagliano, L.; Merlino, A.; Zagari, A.; Mazzarella, L. Proteins 2002, 46, 97.
7. Merlino, A.; Vitagliano, L.; Ceruso, M. A.; Nola, A. D.; Mazzarella, L. Biopolymers 2002, 65, 274.
8. Lai, Y. -T.; Cheng, C. -S.; Liu, Y. -N.; Liu, Y. -J.; Lyu, P. -C. Proteins 2008, 72, 1189.
9. Radkiewicz, J. L.; Brooks, C. L. J Am Chem Soc 2000, 122, 225.
10. Sturtevant, J. M. Proc Natl Acad Sci USA 1977, 74, 2236.
11. Cheng, J. -W.; Lepre, C. A.; Moore, J. M. Biochemistry 1994, 33, 4093.
12. Rischel, C.; Madsen, J. C.; Andersen, K. V.; Poulsen, F. M. Biochemistry 1994, 33, 13997.
13. Hodsdon, M. E.; Cistola, D. P. Biochemistry 1997, 36, 2278.
14. Zidek, L.; Novotny, M. V.; Stone, M. J. Nat Struct Biol 1999, 6, 1118.
15. Fischer, S.; Verma, C. S. Proc Natl Acad Sci USA 1999, 96, 9613.
16. Fischer, S.; Smith, J. C.; Verma, C. S. J Phys Chem B 2001, 105, 8050.
17. Balog, E.; Becker, T.; Oettl, M.; Lechner, R.; Daniel, R.; Finney, J.; Smith, J. C. Phys Rev Lett 2005, 93, 028103.
18. Schmid, F. F. -F.; Meuwly, M. J Mol Biol 2007, 374, 1270.
19. Amadei, A. A.; Linssen, B. M.; Berendsen, J. C. Proteins 1993, 17, 412.
20. Alakent, B.; Doruker, P.; Camurdan, M. C. J Chem Phys 2004, 120, 1072.
21. Alakent, B.; Doruker, P.; Camurdan, M. C. J Chem Phys 2004, 121, 4759.
22. Go, N.; Noguti, T.; Nishikawa, T. Proc Natl Acad Sci USA 1983, 80, 3696.
23. Hayward, S.; Kitao, A.; Go, N. Protein Sci 1994, 3, 936.
24. Hayward, S.; Kitao, A.; Go, N. Proteins 1995, 23, 177.
25. Fierke, C. A.; Johnson, K. A.; Benkovic, S. Biochemistry 1987, 26, 4085.
26. Falzone, C. J.; Wright, P. E.; Benkovic, S. J. Biochemistry 1994, 33, 439.
27. Sawaya, M. R.; Kraut, J. Biochemistry 1997, 36, 586.
28. Cansu, S.; Doruker, P. Biochemistry 2008, 47, 1358.
29. Case, D. A.; Cheatham, T. E., III; Darden, T.; Gohlke, H.; Luo, R.; Merz, K. M., Jr.; Onufriev, A.; Simmerling, C.; Wang, B; Woods, R. J. J Comput Chem 2005, 26, 1668.
30. Duan, Y.; Wu, C.; Chowdhury, S.; Lee, M. C.; Xiong, G.; Zhang, W.; Yang, R.; Cieplak, P.; Luo, R.; Lee, T.; Caldwell, J.; Wang, J; Kollman, P. J Comput Chem 2003, 24, 1999.
31. Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. J Chem Phys 1983, 79, 926.
32. Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. G. J Chem Phys 1995, 103, 8577.
33. Berendsen, H. J. C.; Postma, J. P. M.; van Gunsterena, W. M.; DiNola, A.; Haak, J. R. J Chem Phys 1984, 81, 3684.
34. Ryckaert, J. -P.; Ciccotti, G.; Berendsen, H. J. C. J Comput Phys 1977, 23, 327.
35. Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. Nucleic Acids Res 2000, 28, 235.
36. Brooks, B.; Karplus, M. Proc Natl Acad Sci USA 1983, 80, 6571.
37. Swaminathan, S.; Ichiye, T.; Van Gunsteren, W.; Karplus, M. Biochemistry 1982, 21, 5230.
38. Jackson, J. E. J Qual Technol 1980, 12, 201.
39. Kitao, A.; Hayward, S.; Go, N. Proteins 1988, 33, 496.
40. Umeyama, S. IEEE T Pattern Anal 1991, 13, 376.
41. Brüschweiler, R. J Chem Phys 1995, 102, 3396.

42. Box, G. E. P.; Jenkins, G. M. Time Series Analysis Forecasting And Control; Holden-Day: San Francisco, 1970:Chapters 1–6.
43. Alakent, B.; Camurdan, M. C.; Doruker, P. J Chem Phys 2005, 123, 144911.
44. Hartigan, J. A. Clustering Algorithms; Wiley: USA, 1975:Chapter 4.
45. Cooper, A. Curr Opin Chem Biol 1999, 3, 557.
46. Amadei, A.; De Groot, B. L.; Ceruso, M. -A.; Paci, M.; Di Nola, A.; Berendsen, H. J. Proteins 1999, 35, 283.
47. Verma, C. S.; Caves, L. S. D.; Hubbard, R. E.; Roberts, G. C. K. J Mol Biol 1997, 266, 776.
48. Holm, L.; Sander, C. J Mol Biol 1991, 218, 183.
49. Hayward, S.; Berendsen, H. J. C. Proteins 1998, 30, 144.
50. Hess, B. Phys Rev E 2000, 62, 8438.
51. Box, G.; Kramer, T. Technometrics 1992, 34, 251.
52. Balog, E.; Smith, J. C.; Perahia, D. Phys Chem Chem Phys 2006, 8, 5543.
53. Moritsugu, K.; Smith, J. C. J Phys Chem B 2006, 110, 5807.
54. Kitao, A.; Hirata, F.; Go, N. Chem Phys 1991, 158, 447.
55. Ben-Avraham, D. Phys Rev B 1993, 47, 559.
56. Bahar, I.; Atilgan, A. R.; Erman, B. Folding Design 1997, 2, 173.
57. Atilgan, A. R.; Durell, S. R.; Jernigan, R. L.; Demirel, M. C.; Keskin, O.; Bahar, I. Biophys J 80, 505.
58. Miller, D. W.; Agard, D. A. J Mol Biol 1999, 286, 267.
59. Polshakov, V. I.; Birdsall, B.; Feeney, J. J Mol Biol 2006, 356, 886.
60. Moritsugu, K.; Kidera, A. J Phys Chem B 2004, 108, 3890.
61. van den Bosch, P. P. J.; van der Klauw, A. C. Modeling Identification and Simulation of Dynamical Systems; CRS Press: USA, 1994:Chapter 3.
62. Clarage, J. B.; Romo, T.; Andrews, B. K.; Pettitt, B. M.; Phillips, G. N., Jr. Proc Natl Acad Sci USA 1995, 92, 3288.
63. Boehr, D. D.; McElheny, D.; Dyson, H. J.; Wright, P. E. Proc Natl Acad Sci USA 2010, 107, 1373.