ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/5652474

Conformational Switching upon Phosphorylation: A Predictive Framework Based on Energy Landscape Principles

ARTICLE *in* BIOCHEMISTRY · MARCH 2008

Impact Factor: 3.02 · DOI: 10.1021/bi701350v · Source: PubMed

CITATIONS
29

READS
21

3 AUTHORS, INCLUDING:

Peter Wolynes
Rice University

359 PUBLICATIONS **29,084** CITATIONS

SEE PROFILE

Available from: Peter Wolynes
Retrieved on: 29 January 2016

Conformational Switching upon Phosphorylation: A Predictive Framework Based on Energy Landscape Principles

Joachim Lätzer, Tongye Shen, and Peter G. Wolynes*

Department of Chemistry & Biochemistry, University of California, San Diego, NSF Center for Theoretical Biological Physics, La Jolla, California 92093-0365

Received July 9, 2007; Revised Manuscript Received December 9, 2007

ABSTRACT: We investigate how post-translational phosphorylation modifies the global conformation of a protein by changing its free energy landscape using two test proteins, cystatin and NtrC. We first examine the changes in a free energy landscape caused by phosphorylation using a model containing information about both structural forms. For cystatin the free energy cost is fairly large indicating a low probability of sampling the phosphorylated conformation in a perfectly funneled landscape. The predicted barrier for NtrC conformational transition is several times larger than the barrier for cystatin, indicating that the switch protein NtrC most probably follows a partial unfolding mechanism to move from one basin to the other. Principal component analysis and linear response theory show how the naturally occurring conformational changes in unmodified proteins are captured and stabilized by the change of interaction potential. We also develop a partially guided structure prediction Hamiltonian which is capable of predicting the global structure of a phosphorylated protein using only knowledge of the structure of the unphosphorylated protein or vice versa. This algorithm makes use of a generic transferable long-range residue contact potential along with details of structure short range in sequence. By comparing the results obtained with this guided transferable potential to those from the native-only, perfectly funneled Hamiltonians, we show that the transferable Hamiltonian correctly captures the nature of the global conformational changes induced by phosphorylation and can sample substantially correct structures for the modified protein with high probability.

Protein phosphorylation is one of the most important intracellular control mechanisms (*1*). In both eukaryotic and prokaryotic cells, phosphorylation is a key step in cell cycle control, gene regulation, learning and memory (*2*). Nowadays it is believed that about a third of the proteins in mammalian cells are phosphorylated at one time or another (*3*). Communication in the cell by means of phosphorylation is rapid, reversible and does not require the slow production of new proteins or degradation of existing proteins. Ultimately the activities of proteins that are modified by phosphorylation must be traced to changes in the protein's conformation (*4*−*6*) that are induced by modifying the energy landscape. While native ensembles possess numerous conformational substates, the landscapes of most proteins are highly funnel-like. In many cases, phosphorylation modulates the stability of two near degenerate but structurally distinct conformational ensembles on the landscape allowing the same protein molecule to carry out different activities in the cell at different times. By modulating this near-degenerate landscape, phosphorylation can act as a molecular switch, turning a specific conformation dependent activity on or off by tipping the balance of the population between the two ensembles.

Upon phosphorylation, a phosphate group becomes covalently attached to the side chain of a serine, threonine, tyrosine or histidine residue. Much like the more labile changes due to pH, the change of electric charge in a specific residue through phosphorylation can have several different structural consequences: it can induce local and/or global conformational change between discrete completely folded configurations, or induce order to disorder or disorder to order transitions (*7*). Sometimes the effects of phosphorylation on the structure of the protein appear to be small but further recognition events essential to function, such as binding, can be profoundly affected.

To illustrate how energy landscape ideas can be used to think about phosphorylation and to devise predictive algorithms, we present a theoretical study of how phosphorylation modifies the global (*8*−*10*) rather than local (*11*−*13*) structure of two different proteins, the cysteine proteinase inhibitor cystatin and the receiver domain of the bacterial enhancer-binding protein NtrC[1] (nitrogen regulatory protein C). These two different systems are small enough for detailed theoretical analysis but also have been structurally explored in the laboratory providing thereby the basis for a comparative study to elucidate the generality and specificity of phosphorylation effects.

Cystatins are inhibitors of cysteine proteinases, which destroy proteins by hydrolysis and hence are important in protein degradation (PDB codes 1A67, 1A90) (*14*). Chicken

† Funding was provided by NIH Grant R01 GM044557.

* Corresponding author. E-mail: pwolynes@ucsd.edu. Phone: (858) 822-4825. Fax: (858) 822-4560.

[1] Abbreviations: AMH, associative memory Hamiltonian; LRT, linear response theory; MSD, mean-square-deviation; NtrC, nitrogen regulatory protein C; PC, principal component; PCA, principal component analysis.

10.1021/bi701350v CCC: $40.75 © 2008 American Chemical Society
Published on Web 01/17/2008

cystatin has been structurally characterized in both an unphosphorylated and phosphorylated form. The phosphorylated residue, Ser80, is located in a flexible region of the protein, which is readily accessible both to protein kinases and to phosphatases. Serine phosphorylation sites in many proteins are often found to be flexible or disordered in structural studies. Phosphorylation in intrinsically disordered regions of the protein commonly results in the ordering of the structure in the vicinity of the phosphorylation site (*15*). Unphosphorylated cystatin is a five-stranded β-pleated sheet which is twisted and wrapped partially around a five-turn helix. When cystatin becomes phosphorylated, moderate structural changes occur. The overlay of the mean NMR structures of phosphorylated and unphosphorylated cystatin show an rms deviation between the structures of 2.7 Å. Cystatin thus serves as a paradigm for a system having minimal structural change induced through phosphorylation in a flexible loop region.

A more dramatic change upon phosphorylation in terms of structure occurs in another well characterized system, the receiver domain of NtrC. The receiver domain of NtrC is a conformational switch found in a bacterial "two-component'' regulatory system (PDB codes 1DC7, 1DC8) (*16*). Upon phosphorylation two β-strands as well as two α-helices are displaced away from the phosphorylation site and additionally one helix is rotated axially. The overlay of the average NMR structures of the unphosphorylated and phosphorylated conformation of NtrC shows larger rms deviation between the structures of about 3.3 Å. The amplitude of the change is thus slightly larger than for cystatin. NtrC has been regarded as a model for a conformational switch (*17*), in which a "large'' conformational change is induced upon phosphorylation. Clearly larger proteins can exhibit still larger changes in an rms sense, owing to a greater lever arm for hinge motion in them.

The aim of the current study is to elucidate how phosphorylation causes these observed changes in protein conformations. First we examine the free energy profiles that would be obtained by assuming an ideal landscape having as little frustration as possible. This landscape for the phosphoprotein is constructed by utilizing the information about the structures of both phosphorylated and unphosphorylated native forms. Such a model yields the free energy difference of the forms that would be expected if only the native contacts were to contribute to the energetics. Since the conformations and hence the contact maps of the unphosphorylated and the phosphorylated proteins in our study are already known from experiments, we can construct such a structure based Hamiltonian having native-only interactions for molecular dynamics simulations to obtain conformations and energies of the proteins along the reaction coordinate. This is a "vanilla'' Hamiltonian because it is topology based, not singling out any interactions as especially significant. This model treats the two different sets of input native contacts, those for the unphosphorylated conformation and those for the phosphorylated one, as independent. We can more directly extract changes in the free energy profile using the free energy perturbation method. Next, a principal component analysis of the contact maps of the simulated ensembles allows us to find the dominant components of the phosphorylation induced change and to visualize the effect that phosphorylation has on a residue−residue contact map. The contact maps of the test proteins in the unphosphorylated and phosphorylated forms show that many of the contacts formed by the phospho residues for the test proteins are preserved, suggesting the effect of phosphorylation primarily lies in the long-range forces. This observation allows us to address a rather practical issue: Instead of needing structural information on both forms, can one predict the likely conformational changes that should occur when only one structural form is known? For example, given structural information only about the unphosphorylated protein and the sequence information of which particular residues are susceptible to phosphorylation, can one predict the dominant conformation of the phosphorylated protein? For such predictions, obviously perfect funnel, native topology based models will not suffice. Since long-range interactions are expected to be dominant however, we can construct a new guided structure prediction Hamiltonian by using local structural information known from the unphosphorylated protein for residue interactions separated by a few residues (12 in this case), but use a transferable structure prediction Hamiltonian (AMH) (*18, 19*) having a heterogeneous through space potential for residues that are more than 12 residues apart in sequence. The transferable long-range potential while transferable has been shown to yield a reasonably funneled potential which has been optimized based on a large set of generic protein structures to successfully predict the folded state of proteins of size up to 180 residues. Its predictive power has been well documented (*20*). Additionally it is possible to construct a new potential in this format to evaluate the interactions of the phosphorylated residues based on the same form.

To obtain the Hamiltonian for phosphorylated proteins from that which has been optimized for normal, unphosphorylated amino acids, we earlier postulated that we can treat the interactions involving the phosphorylated residue as those of a "supercharged'' glutamic acid residue (*21*). The energetic interactions of the phosphorylated residue with other residues are replaced with enhanced interactions of the type ordinarily used for a glutamic acid residue with the corresponding residues (*21*). Since the energy landscape of the unphosphorylated protein is known and the contact maps of the test proteins indicate there is a considerable overlap of contacts between the unphosphorylated and phosphorylated conformations, we preserve the native focused associative memory terms biased toward the assumed known unphosphorylated structure for residues that are less than 12 residues apart in sequence space but use the transferable potential with a "supercharged glutamate'' for the more distant interactions. The Hamiltonian we have constructed in this way equips us with an energy function that should reliably mimic the local structure of the unphosphorylated protein, but that nevertheless plausibly treats the effects of the long-range forces on the conformation of the protein. We show this Hamiltonian correctly predicts many features of the conformational changes observed in the phosphorylated protein. To document that this can be done, we set up simulations with different strengths of the charge interactions for the phosphorylated protein, and then we project the conformations obtained in these simulations onto the first two principal components obtained earlier using the "vanilla'' native-structure-based Hamiltonian. We also show more directly

that structures rather close to the NMR structures can also be sampled.

METHODS

In order to explore the issues raised above, we studied four Hamiltonians based on the native configurations of the test proteins, $\mathcal{H}_u$, $\mathcal{H}_p$, $\mathcal{H}^*_u$, and $\mathcal{H}^*_p$. We show how to construct the native-based Hamiltonians $\mathcal{H}_u$, $\mathcal{H}_p$ in the first subsection. These two Hamiltonians are based on the information of the experimentally determined native structures of the unphosphorylated or phosphorylated form of the proteins. Note that throughout the current study, we use the subscripts *p* and *u* to indicate the phosphorylated or the unphosphorylated form respectively. We then describe how to obtain the free energy profiles from the conformations sampled with $\mathcal{H}_u$ and $\mathcal{H}_p$ and describe a principal component analysis based on the contact maps of these conformations.

Finally we describe the construction of structure prediction Hamiltonians $\mathcal{H}^*_u$ and $\mathcal{H}^*_p$, *both* of which are based on transferable interactions using the long range interaction parameters optimized for generic structure prediction but that use information about the native conformation of the unphosphorylated form to encode the short- and intermediate-range interactions. Note that neither $\mathcal{H}^*_u$ or $\mathcal{H}^*_p$ contains any experimental information of long-range interactions found in the unphosphorylated form; neither $\mathcal{H}^*_u$ or $\mathcal{H}^*_p$ directly makes use of any (short-, intermediate-, or long-range) experimental information on the *phosphorylated* form at all.

We also detail how we define various physical quantities for monitoring structural ensembles, such as order parameters and configurational free energy, which we adopt to analyze the results of all simulations based on these four Hamiltonians.

1. Native-Structure-Based Simulations. Simulations of the folding dynamics of cystatin and NtrC were performed with an off-lattice native-structure-based potential. The Hamiltonian used in this study contains a basic backbone Hamiltonian and a contact potential

$$\mathcal{H}_{u/p} = \mathcal{H}_{bb} + \mathcal{H}_{c,u/p} \tag{1}$$

and depends on the locations of the C^α, C^β and oxygen atoms. The index *u*/*p* is a simplified notation for the two cases, namely *u* or *p*. The remaining backbone atom positions can be calculated assuming ideal backbone geometry. The backbone potential H_{bb} constrains the backbone to have chemically and physically acceptable conformations (*22*). The backbone potential is given by

$$\mathcal{H}_{bb} = \lambda_{\psi\phi}\mathcal{H}_{\psi\phi} + \lambda_{\chi}\mathcal{H}_{\chi} + \lambda_{ex}\mathcal{H}_{ex} + \lambda_{harmonic}\mathcal{H}_{harmonic} \tag{2}$$

The Ramachandran potential $\mathcal{H}_{\psi\phi}$ provides a good fit of the backbone torsional angles based on the statistics of protein structural database. The chirality potential $\mathcal{H}_{\chi}$ biases the protein chain into the L-amino acid configuration. The algorithm SHAKE constraints for the heavy backbone atoms along with three quadratic potentials provide for backbone rigidity and planarity. To complete the picture of stereochemically allowed protein backbones, an excluded volume potential is applied to the oxygen and carbon atoms of residue *i* and *j*. This potential applies when the heavy atoms approach within 3.5 Å for residues close in sequence space such that $(j - i) < 5$, and 4.5 Å for $(j - i) \geq 5$. The λ-terms scale the interactions of the individual backbone potentials.

The contact term $\mathcal{H}_c = \mathcal{H}_{c,S} + \mathcal{H}_{c,M} + \mathcal{H}_{c,L}$ is an associative memory term (*23*). Through its guidance, the free energy will reach a minimum at the basin of the given native PDB structure. Since there are several structures of cystatin deposited in the PDB, all these structures were used as memory terms for the simulation. The functional form of the contact term is given by

$$\mathcal{H}_{c,u/p} = -\epsilon \sum_{i \leq j-3} \gamma[x(|i-j|)] \exp\left[-\frac{(\mathbf{r}_{ij} - \mathbf{r}_{ij}^{N\,at,u/p})^2}{2\sigma_{ij}^{\,2}}\right] \tag{3}$$

The sum runs over all carbon atom pairs (C^α–C^α, C^α–C^β, C^β–C^α, C^β–C^β) having a sequence separation of at least three residues. The functional form of the interactions of the carbon atoms in this potential are Gaussian centered at the native distance $r_{ij}^{N\,at}$ and with a width of $\sigma_{ij} = |i-j|^{0.15}$Å. The $\mathcal{H}_c$ potential depends on the sequence separation $|i - j|$ of the residues *i* and *j*. We divide the energy into three different proximity classes $x(|i - j|)$: short range (S) for $|i - j| < 5$, medium range (M) for $5 \leq |i - j| \leq 12$ and long range (L) for $|i - j| > 12$. The $\gamma[x(|i - j|)]$-terms are weighted such that the energies in each proximity class $x(|i - j|)$ are equal to each other. Also the energies of any contact in each proximity class are equal for all contacts formed. The total energy of the Hamiltonian is scaled to be $4N$, where N is the number of residues of the protein. The unit of energy can then be denoted as ϵ and is defined in terms of its native state energy coming from the contact term $\mathcal{H}_c$ only,

$$\epsilon = \frac{\langle H_c \rangle}{4N} \tag{4}$$

The simulation protocol is as follows: For each protein twenty constant temperature runs were performed with the structure based Hamiltonian. The constant temperature runs sampled 800 independent structures each spaced at intervals at about 1 μs corresponding to a trajectory of about 1 ms in physical time. A total of $16000 \times 2 \times 2 = 64000$ structures were obtained for various temperatures for the unphosphorylated protein as well as for the phosphorylated protein. The key thermodynamic quantity desired from the simulations is the free energy as a function of reaction coordinate *Q* and temperature. The normalized collective coordinate *Q* measures the similarity of two conformations A and B to each other.

$$Q = \frac{2}{(N-1)(N-2)} \sum_{i<j-1} \exp\left[-\frac{(r_{ij}^A - r_{ij}^B)^2}{2\sigma_{ij}^{\,2}}\right] \tag{5}$$

2. Free Energy Perturbation Method. We directly examine how phosphorylation changes the free energy profiles. We start by analyzing the sampling snapshots obtained in simulation with each of the two Hamiltonians. After projecting the ensembles to the desired collective coordinates $r = \{r_1, r_2, ...\}$, the probability distribution $\rho(r) = N(r)/N_{tot}$ is computed for a total of N_{tot} snapshots. We can then derive

straightforwardly the free energy profile $F(r) = -k_BT \ln(\rho(r)/\rho_0)$, where ρ_0 is a uniform distribution, for the unphosphorylated and phosphorylated conformations. We are interested in the difference of the two free energies, so we subtract these to obtain the difference $\Delta F(r) = F_p(r) - F_u(r)$.

In the current case we use two different folding order parameters Q_u and Q_p as the collective coordinates, i.e., $r = (Q_u,Q_p)$. We assign to each snapshot two numbers, the order parameters Q_u and Q_p, which measure how similar an individual snapshot obtained in the MD simulations is to the native structure of the unphosphorylated or the phosphorylated conformations respectively. Simulations with the native-structure-based Hamiltonians bias the sampled conformations strongly toward the native structure. Performing a simulation with one of the two Hamiltonians, say $\mathscr{H}_u$, results in greater sampling of structures with high Q_u but sparse sampling of structures with high Q_p.

Instead of using a brute force approach of performing a large amount of simulations to ensure acceptable sampling of the 2D reaction coordinate space, we use the free energy perturbation method (*24*) to obtain the free energy difference directly. Thus to calculate the free energy difference ΔF from the sampling of the unphosphorylated Hamiltonian $\mathscr{H}_u$, we not only project the sampled conformations to the collective Q coordinates but also record, for each conformation, what the energy $E_u = \langle\mathscr{H}_u\rangle$ of the unphosphorylated system is and also what the energy $E_p = \langle\mathscr{H}_p\rangle$ of a phosphorylated system with the *same* conformation would be. We then perform the statistics on the raw moments of the energy difference $\langle\Delta E^k\rangle(r) = \langle(E_p - E_u)^k\rangle(r)$. The free energy difference of the two systems is then simply given by the cumulant expansion equation, i.e.,

$$\Delta F(Q_u,Q_p) = -\sum_{j=1}[(-\beta)^j/j!]C_j(Q_u,Q_p) \qquad (6)$$

Here C_j is the jth order of the expansion. We have $C_1 = \langle\Delta E\rangle$, $C_2 = \langle\Delta E^2\rangle - \langle\Delta E\rangle^2$, etc.

3. Contact Map Principal Component Analysis. We also use a principal component analysis (PCA) based on contact maps to visualize the conformational changes induced by phosphorylation. The more commonly used principal component analysis based on the diagonalization of the Cartesian coordinates is less useful for our purposes because the change in the energy is only weakly related to the changes in the linear Cartesian distances. This mismatch is due to the fact that in phosphorylation the large conformational changes are generally of a magnitude beyond the simple vibrational-like fluctuations of the Cartesian coordinates. To capture properly the conformational changes, it is necessary to employ a set of detailed, site specific, and structure based reaction coordinates that do correlate with the energy. The global order parameters Q_u or Q_p do not suffice for the detailed description. We select a set of coarse-grained yet local-information-revealing degrees of freedom encoded in the contact map. This is the simplest site specific measure properly capturing the structure of a conformation while relating directly to the energy. A contact between residues i and j is considered to be formed (given the value of 1 as opposed to 0 when no contact is formed) when the distance of the respective C^β atoms is less than 6.5 Å. For each snapshot obtained in the molecular dynamics we compute the contact map. The contact principal component analysis (*25*) reflects the correlations between different contact forming events. The covariance matrix to be diagonalized is not based on the linear Cartesian coordinates but rather on a contact map correlation function

$$C_{i,j,k,l} = \langle(m_{ij} - \langle m_{ij}\rangle)(m_{kl} - \langle m_{kl}\rangle)\rangle \qquad (7)$$

This "hypermatrix'' encodes how an instance in which residues i and j form a contact correlates to an instance where residues k and l form a contact. To further facilitate the analysis, we coarse-grained the contacts by grouping neighboring residues into groups of four residues, i.e., a coarse-grained contract matrix is calculated for each snapshot, with each of the independent elements being either 0 or 1. The coarse-grained contacts are reduced in number to $27 \times (27 - 1)/2 = 378$ and $31 \times (31 - 1)/2 = 465$ for cystatin and NtrC respectively. The resulting reduced covariance matrices of dimension 378×378 and 465×465 are diagonalized, and the eigenvalues are calculated. The two most dominant principal components (PC) are plotted.

4. Linear Response Theory (LRT). As an alternative to the detailed sampling of the predictive Hamiltonian in the next subsection, we can use the linear response theory to see how phosphorylation should induce conformational changes. Linear response theory suggests that the magnitude of the conformational changes is a convolution of the strength of the sequence specific perturbation times the susceptibility of the corresponding degrees of freedom to make such changes (*26*, *27*). Statistical thermodynamics shows the coefficient of the response of a system under small external change is also linearly related to the fluctuations of the system sampled at equilibrium. The most commonly known manifestation of this relation explains how the heat capacity, a measurement of how energy changes with the temperature change of a system, is related to energy fluctuations.

In our case, the linear response theory describes the changes of the contact map using a relation of the form

$$\langle\delta q_{i,j}\rangle = \sum_{k,l} C_{i,j,k,l}\,\langle\delta V_{k,l}\rangle \qquad (8)$$

where $\delta V_{k,l}$ is the matrix of contact energy change upon phosphorylation. The details of δV will be spelled out in detail in the next subsection. Nevertheless it is easy to see that δV is a very local property in the contact representation. For example, say residue 7 is the only residue that undergoes phosphorylation, we will then only have nonzero contributions of δV for the elements $\delta V_{k,l}$ if $k = 7$ or $l = 7$, otherwise $\delta V_{k,l} = 0$. By bridge in with the hypermatrix $C_{i,j,k,l}$, we can see how the changes of contact energy between the pair $i-j$ are correlated with the changes of contact probability between the pair $k-l$ at equilibrium. Linear response analysis yields the change in probability of forming a certain pair $i-j$ when all the input contact energies change. Since δV is very local, i.e., is an extremely sparse matrix, it follows that the structural responses are primarily a combination of the largest eigenvectors of the diagonalization of the hypermatrix C (the top PCs). The dominance of these modes reflects the fact that those eigenvectors have largest amplitude of fluctuation. The linear response theory is an efficient method to give a quick estimate of the changes caused by a perturbation. It is

more accurate for systems that undergo small changes than for systems that undergo complicated, more involved changes.

5. Modeling Tertiary Structure Effects of Phosphorylation. Can one predict the conformation of the phosphorylated protein given knowledge of the folding landscape of the unphosphorylated protein only and the changes in the modifiable residues? As a first step to answer the prediction question, we developed a set of Hamiltonians $\mathcal{H}^*$ based on the information of the unphosphorylated form alone. We use the superscript $*$ to denote the energy functions that are transferable to distinguish the two sets. We first compare the difference of the ensembles generated by $\mathcal{H}_p$ and $\mathcal{H}_u$ and the difference of the ensembles generated by $\mathcal{H}^*_p$ and $\mathcal{H}^*_u$. We thus constructed a specific Hamiltonian constructed in the following form:

$$\mathcal{H}^*_{u/p} = \mathcal{H}^*_{c,L,u/p} + \mathcal{H}_{c,S+M,u} + \mathcal{H}_{bb} \quad (9)$$

The only difference between $\mathcal{H}^*_{u/p}$ and $\mathcal{H}_u$ lies in the long-range energy terms. All three Hamiltonians share the same backbone and the same short and intermediate contact terms with each other. Here $\mathcal{H}_{c,S+M,u}$ is given by eq 3 and summed only over residues that are separated by twelve or fewer residues in sequence space. This term biases the local secondary structure of the protein by having only the native interactions of one of the forms and hence yields largely native secondary structure. The tertiary structure of the protein follows thus from the contact energy term. This contact energy term arises from an optimized energy function used previously for protein structure prediction. The details may be found in (*19*) and references therein. A 4-letter code is utilized and the specific amino acids in each category are denoted as hydrophilic (Ala, Gly, Pro, Ser, Thr), hydrophobic (Cys, Ile, Leu, Met, Phe, Trp, Tyr, Val), acidic (Asn, Asp, Gln, Glu) and basic (Arg, His, Lys). The energy contributions of the contact potential to the total potential are given by a three-well potential.

$$\mathcal{H}^*_{c,L,u/p} = -\epsilon^* \sum_{i<j-12} \sum_{k=1}^{3} \gamma^*(P_i,P_j,k)\, c_k(N) \times U[r_{min}(k),r_{max}(k),r_{ij}] \quad (10)$$

Here k is a function of the spatial distance r_{ij} of residues i and j and c_k is found from fitting the number of contacts of the protein in each of the regions of k as a function of sequence length of the target protein. The interactions are weighted by the interacting amino acids of class P_i and P_j and their spatial distance. The parameters γ^* have been optimized based on the principle of minimal frustration. It is critical to note that γ^* is a function of residue chemistry, thus γ^*_u and γ^*_p have different values. More specifically, γ^*_u was derived from a structural database of ordinary, unphosphorylated proteins following the training procedure for the parameters based on the quantitative form of the minimal frustration principle (*22*). The training maximizes the energy gap over the variance. This quantity is a measure of how funneled the landscape is toward a properly folded structure as compared to a random ensemble of molten globule structures. The procedure for deriving the parameters has been described in greater detail by Hardin et al. The contact function U controls the shape and sharpness of the multiwell potential (*22*). It is important to stress that this term is heterogeneous but generic and transferable. As for γ^*_p, we have modeled the influence of the phosphorylation of an amino acid by substituting for the phosphorylated residue a supercharged glutamic acid residue. This strategy was put forward in previous studies of phosphorylation of NFAT where the structure was entirely unknown (*21*). An analogous experimental approach based on the analogy between phosphoserine and glutamate has also been demonstrated to work in several cases, notably in studies on the dematin headpiece (*28*) and tumor suppressor protein p53 (*29*). These studies show that the Ser-to-Glu mutant closely mimics the conformation of the phosphorylated protein. The details of the implementation of the hypercharged residue and its interactions with other residues as well as robustness and caveats have already been described by Shen et al. (*21*).

As $\mathcal{H}^*_u$, $\mathcal{H}^*_p$ and $\mathcal{H}_u$ share the same values for all other energy terms, it would seem to be extremely demanding to try to predict the exact changes based on this generic long-range term alone. Still we will present quite a successful demonstration of the importance of the generic long-range potential in predicting the phosphorylated conformation. The trends of conformational changes generated by $\mathcal{H}^*_p$ observed in the simulations are consistent with the trends generated by $\mathcal{H}^*_u$ and thus by experiments. Constant temperature MD simulations with the Hamiltonian $\mathcal{H}^*_{c,L}$ were performed to predict the structure of the phosphorylated protein. In these simulations the starting structure was fixed to be the average NMR structure of the unphosphorylated protein. Following this a total of $16000 \times 2 = 32000$ independent structures were sampled.

RESULTS FOR THE NATIVE-STRUCTURE-BASED HAMILTONIANS

1. Free Energy Landscape of Phosphorylated Proteins. To sensibly study global effects of phosphorylation using coarse-grained models, the contact maps of the unphosphorylated and phosphorylated forms of the test proteins must be different, that is, sufficiently large to be reflected in the contact maps of the test proteins. The contact maps of the unphosphorylated and phosphorylated conformations of cystatin and NtrC are shown in Figure 1. The important conformational changes induced by phosphorylation of cystatin do indeed present themselves in the contact map. Phosphorylation however introduces rather minor perturbations to the cystatin system. The contact map of NtrC shows more substantial changes upon phosphorylation. The contacts of the phospho residue in both the unphosphorylated and phosphorylated conformations are identical, but phosphorylation apparently introduced long-range effects that led to the global conformational change of NtrC.

Molecular dynamics simulations with the native-structure-based Hamiltonians were performed to obtain adequate sampling of the conformations of cystatin and NtrC in their unphosphorylated and phosphorylated conformations. First, snapshots of MD simulations were sampled with the unphosphorylated native-structure-based Hamiltonian, $\mathcal{H}_u$. For each snapshot, the energy E_u and the order parameter Q_u, which measures similarity to the average structure of the unphosphorylated conformation, were calculated. The probability distribution ρ was computed. This allows calculation of the free energy, $F(r) = -k_BT \ln(\rho(r)/\rho_0)$. For the same

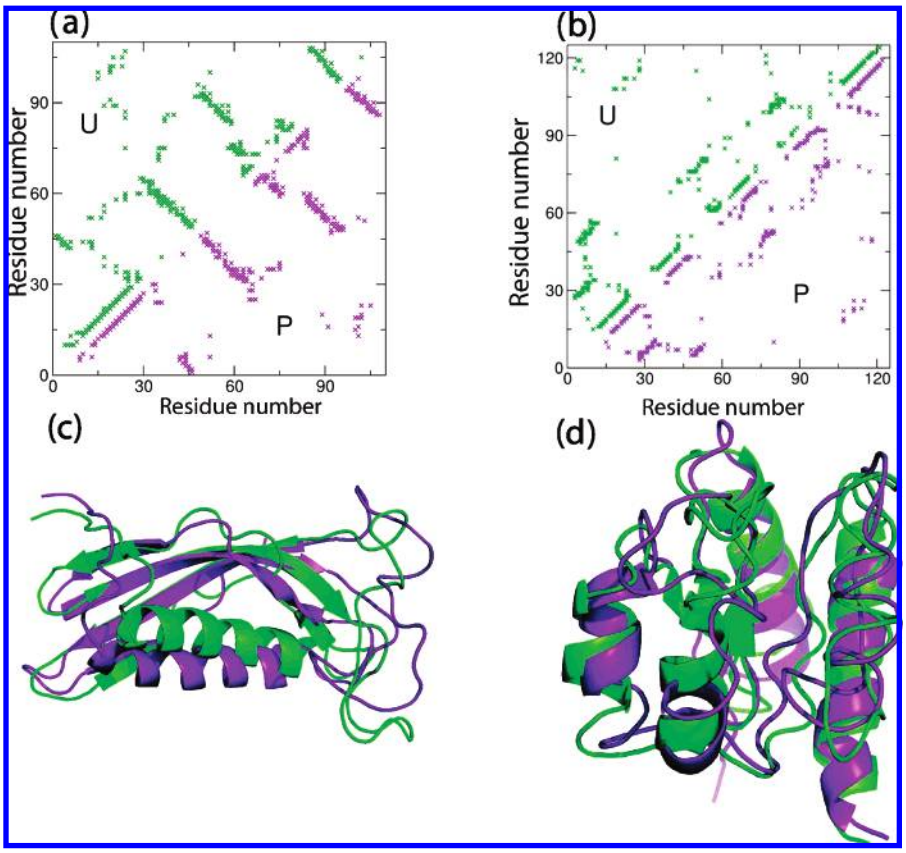

FIGURE 1: Contact maps of (a) cystatin and (b) NtrC and the corresponding structures shown in (c) and (d). The average contact maps for the unphosphorylated conformations are shown in the upper triangle, while the contacts of the phosphorylated protein forms are projected on the lower triangle.

snapshots obtained with $\mathcal{H}_u$, the energy E_p, which can be obtained from the Hamiltonian of the phosphorylated conformation, $\mathcal{H}_p$, and the order parameter Q_p were computed. The 2D free energy profiles of unphosphorylated cystatin and NtrC are plotted in Figures 2 and 3. The set of (E_u,Q_u) and (E_p,Q_p) found for snapshots at various Q_u and Q_p was used to obtain the free energy difference $\Delta F(r) = F_p(r) - F_u(r)$ via the cumulant expansion equation.

The gradient of $\Delta F(r)$ is also plotted in Figures 2 and 3 and is indicated by the arrows on the free energy landscape at each position along the folding order parameter. The lengths of the arrows indicate the relative magnitude and direction of the change of $\Delta F(r)$. Also the same procedure is applied to conformations sampled in molecular dynamics runs with the $\mathcal{H}_p$ Hamiltonian as energy function. The results are plotted in Figures 2 and 3.

The free energy profile for cystatin at a simulation temperature close to the folding temperature of $T = 1.0$ shows a simple two-state folding process with an unfolded and a folded basin (Figure 2) separated by a barrier of about $4k_BT$. The coordinates in Q_u,Q_p of the two free energy minima for the unphosphorylated protein are given by (0.29,0.25) for the unfolded basin and (0.64,0.52) for the folded basin. The free energy minimum for the folded state of the phosphorylated cystatin is located at (0.49,0.62). The gradient of the free energy difference $\Delta F(r)$ is also shown as a vector that gives a good indication at each value of the reaction coordinate, how phosphorylation effects the profile. In the phase space region of $Q_u \leq 0.5$ the arrows point directly into the direction of the phosphorylated protein. This is due to the fact that, before reaching the transition state, the two forms of the unphosphorylated and phosphorylated protein can easily interchange. Even after crossing the transition state, the direction of the gradient of both forms is almost the same as before with the difference that most arrows do point slightly in the direction of lower Q_u, the unfolding direction. Figure 2 shows the free energy plot for sampling of phosphorylated conformations with $\mathcal{H}_p$. The resultant 2D free energy landscape was similar to the landscape obtained with $\mathcal{H}_u$ and using the cumulant expansion method to determine $\Delta F(r)$. Principal component analysis was performed and the conformations were projected onto the first two dominant principal components as shown in Figure 2. For every projected snapshot it is known how folded the structure is and also if the snapshot stems from a simulation of the unphosphorylated or phosphorylated protein. The principal components therefore correspond to folding and phosphorylation, and we can name them the folding principal component PC^{fold} and the phosphorylation principal component PC^{phos}. PC^{fold} measures the general folding order with more negative PC^{fold} indicating a more folded set of structures. PC^{phos} measures how much a conformation is similar to the phosphorylated conformation, that is, the negative direction corresponds to the direction of conformational changes that occur upon phosphorylation. Projection of the changes of PC^{phos} onto a contact map allows inspection of phosphorylation induced contact changes. The PC^{phos} contact map shows the dominating contact changes upon phosphorylation in blue, while contacts dominating in the unphosphorylated form show up in red. Direct comparison of the structural changes of the simulated ensembles (Figure 2d) to the changes observed in the contact map

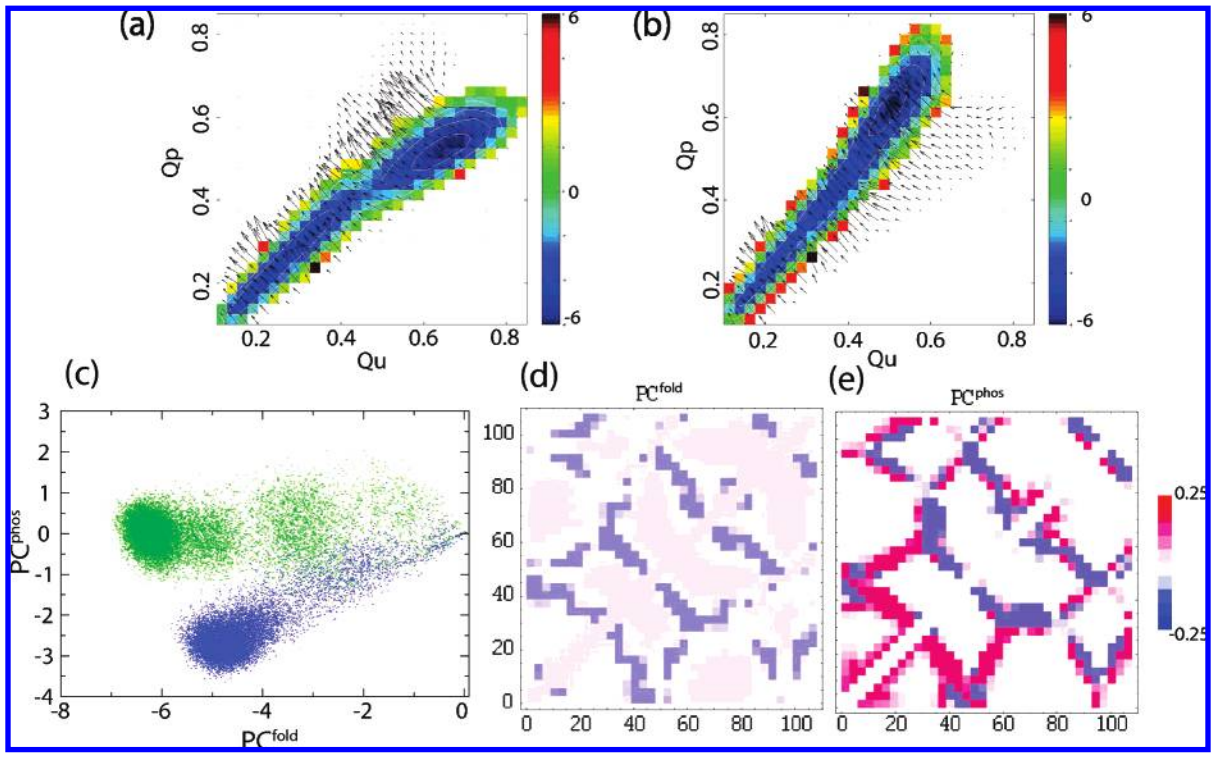

FIGURE 2: Free energy landscapes of cystatin folding for the unphosphorylated form (a) and the phosphorylated form (b). The white contour lines are drawn to facilitate observation of the native and unfolded basins in the free energy landscape. Arrows indicate the gradient of the free energy landscape pointing in the direction of phosphorylation and scaled in size to representable values. Snapshots of the conformations of unphosphorylated cystatin (green) and the phosphorylated cystatin (purple) projected along the first two dominant principal components in (c). The largest two principal components shown in the contact map form (d, e).

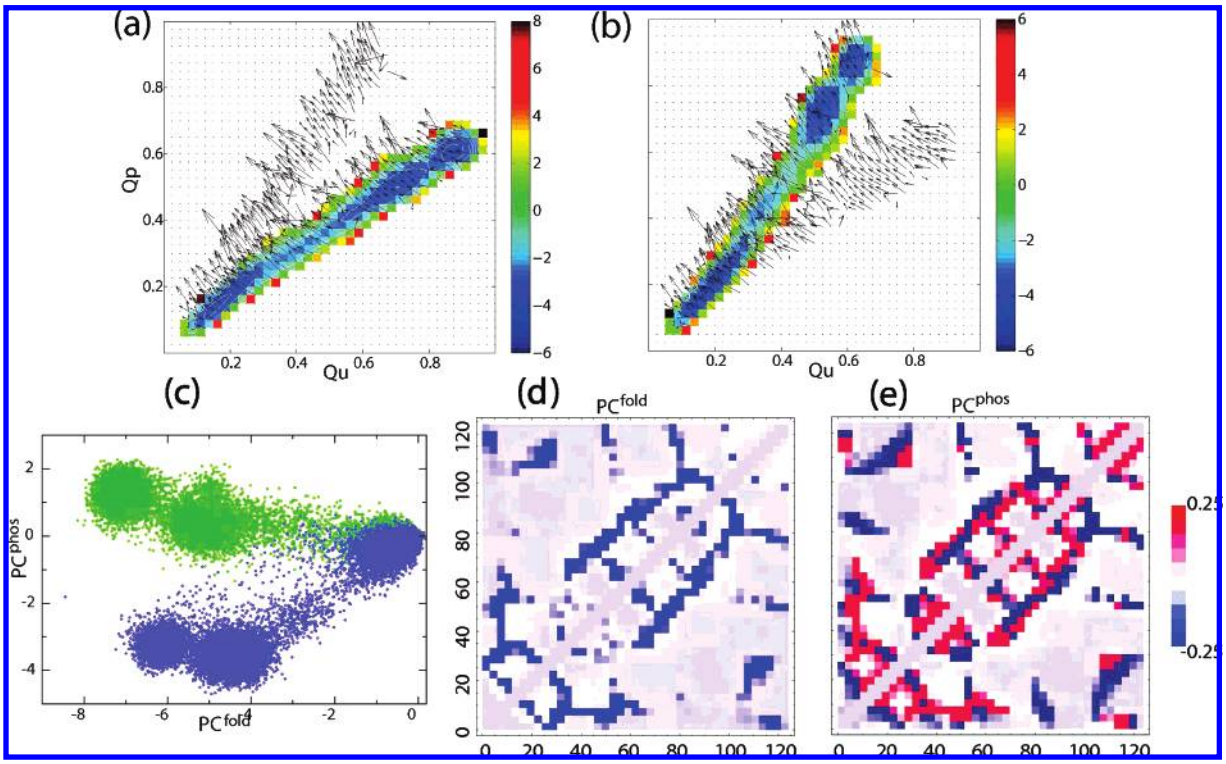

FIGURE 3: Free energy landscapes of NtrC folding for the unphosphorylated form (a) and the phosphorylated form (b). Arrows and contour lines are drawn for better visualization. Snapshots of the unphosphorylated NtrC (green) and the phosphorylated NtrC (purple) projected along the first two dominant principal components in (c). The largest two principal components are shown in the contact map form (d, e).

obtained from the pdb native structures of the unphosphorylated and phosphorylated form (Figure 1a) show excellent agreement, i.e., contacts that are exclusively formed in the unphosphorylated form show up as red while contacts that are solely formed upon phosphorylation show up in blue.

Three free energy minima are found in the free energy plot of unphosphorylated NtrC at temperature $T = 1.0$ (Figure

3). This suggests that the unphosphorylated NtrC is not a two-state folder but has a well-ordered intermediate at coordinates in Q_u,Q_p given by (0.7,0.5). The native basin is located at (0.87,0.6), and the unfolded basin is at (0.17,0.16). The gradient of the free energy difference $\Delta F(r)$ is again plotted using arrows, that indicate the direction and magnitude of the change in $\Delta F(r)$ upon phosphorylation. The arrows show a largest gradient in the intermediate state, which would suggest that transitions from the unphosphorylated conformation to the phosphorylated conformations of NtrC are preferred in the intermediate states of folding. The free energy profile obtained from $\mathcal{H}_p$ for the phosphorylated NtrC is also shown (Figure 3). The folding is also 3-state with three main free energy minima. Principal component analysis was performed on the snapshots obtained in the molecular dynamics simulations with Hamiltonians $\mathcal{H}_u$ and $\mathcal{H}_p$ (Figure 3). It is apparent from the figure that the 3-state folding behavior is well captured by the principal component analysis. The first two components are by themselves very useful in capturing the folding and the effects of phosphorylation respectively. We identify the principal component PC^{fold}, which provides a good indication of the degree of the folding order, where a more negative PC^{fold} indicates a more folded set of conformations. PC^{phos} serves to distinguish the unphosphorylated ensemble from the phosphorylated ensemble. Projection of the first two principal components of snapshots is shown in Figure 3. The agreement with experiment is great, again. Figure 3 proves useful in identifying the trends of contact changes upon phosphorylation. We note that, for a 3-state folder, the third principal component might also be important. Plots of combinations of any two of the first three components show 3-state behavior, however first two principal components do distinguish the global folding and phosphorylation best.

2. Changes in Free Energy Profiles between Unphosphorylated and Phosphorylated Protein Conformations. In vivo, proteins that become phosphorylated can have two sensibly different average conformations as revealed by X-ray crystallography or NMR despite the two forms having obviously almost identical sequences (except for the phospho residues, the two sequences are identical). Normally sequences with high sequence similarity adopt the same fold (*30*). Thus it may seem obvious to assume that in fact the unphosphorylated protein itself can assume both conformations, the unphosphorylated conformation and the phosphorylated conformation. However, for phosphorylation to crisply act as a molecular switch, the two conformations should be separated by a high barrier such that the unphosphorylated protein will not likely spontaneously adopt the incorrect structure and hence function of the phosphorylated protein. It is natural then to ask how difficult is it for the unphosphorylated protein to change from the unphosphorylated basin to the phosphorylated basin. Nature achieves this basin change by an enzymatic reaction that adds a phosphate group to the residue susceptible for phosphorylation. If the energy landscape were perfectly funneled with only a single set of native contacts (as for $\mathcal{H}_u$ and $\mathcal{H}_p$) (*31*), the free energy difference between the basins would be large if the two forms were very different.

In this study the sampling was performed with two different Hamiltonians. To understand the free energy profile for motion between the native (unphosphorylated and phosphorylated) basins, we use a simple approach to determine the barrier location and barrier height. We estimate an effective barrier height by finding the minimum of the intersection of the two basins found in the free energy profiles. A further simplification is made assuming an isotropic, harmonic basin shape. The free energy profile around a basin with minimum position (Q_1,Q_2) is assumed to be of the form of $F(Q_u,Q_p) = (a/2)[(Q_u - Q_1)^2 + (Q_p - Q_2)^2] + F_0$. We study the profile along the reaction coordinates that link two basins (Q_1^u,Q_2^u) and (Q_1^p,Q_2^p) with a simple straight line. Without loss of generality, we assume the narrower of the two basins is at the origin, and the other basin is at distance $b = [(Q_1^u - Q_1^p)^2 + (Q_2^u - Q_2^p)^2]^{1/2}$. Their minima are at 0 and $c = \Delta F$ respectively. Along this one-dimensional coordinate we have $F_1(r) = (a_1/2)r^2$ and $F_2(r) = (a_2/2)(r - b)^2 + c$ under the condition $a_1 \geq a_2$. As shown in Figure 4, the intercept occurs at

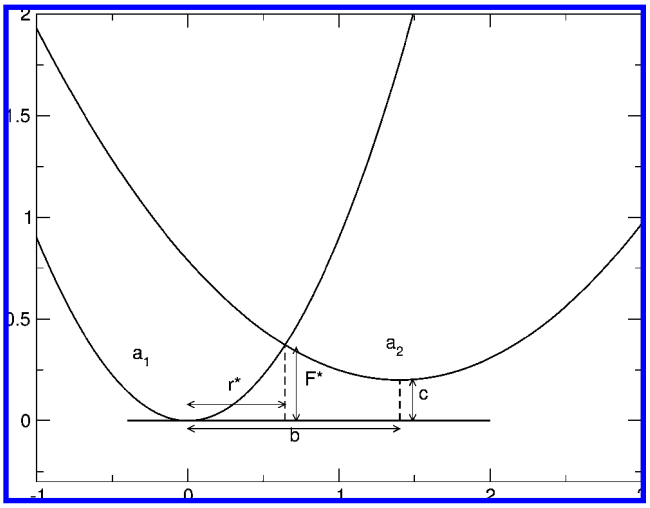

FIGURE 4: The illustration of free energy barrier estimation.

$$r^{\#} = \frac{-a_2 b + [a_1 a_2 b^2 + 2c(a_1 - a_2)]^{1/2}}{a_1 - a_2}$$

The barrier height is then given by $F^{\#} = (a_1/2)r^{\#2}$. If $a_1 = a_2 = a$, then we can compute $r^{\#} = b/2 + c/(ab)$. For the case of cystatin, we found that at $T = 1$, $Q^u = (0.64,0.52)$ and $Q^p = (0.49,0.62)$, we have $b = 0.57$, a rough fit gives $a = 500$ and $c = 0.01$. As a result we found that the barrier height of the free energy is $F^{\#} = 20$ for cystatin. Similarly we find at $T = 0.8T^{room}$, $Q^u = (0.87,0.6)$ and $Q^p = (0.52,0.72)$, $c = 0.5$, $b = 1.17$, and $a = 600$, we found $F^{\#} = 90$ for NtrC. The unit of barrier height is given by $k_BT \sim 0.6$ kcal/mol. Note that both numbers seem rather high. As explained by Miyashita et al. (*32*) the local quadratic approximations are first of all quite rough and should only lead to an approximate barrier with the right order of magnitude. In reality, the barrier is much lower, because the transition state is not necessarily located on the straight line connecting the unphosphorylated basin with the phosphorylated basin. The height of the barrier should be interpreted as follows: In the context of a perfectly funneled landscape to a single minimum, the barrier located on the direct route between the unphosphorylated basin and phosphorylated basin of cystatin would be so large as to prevent an equilibrium of both conformations at the same time. We see this allows

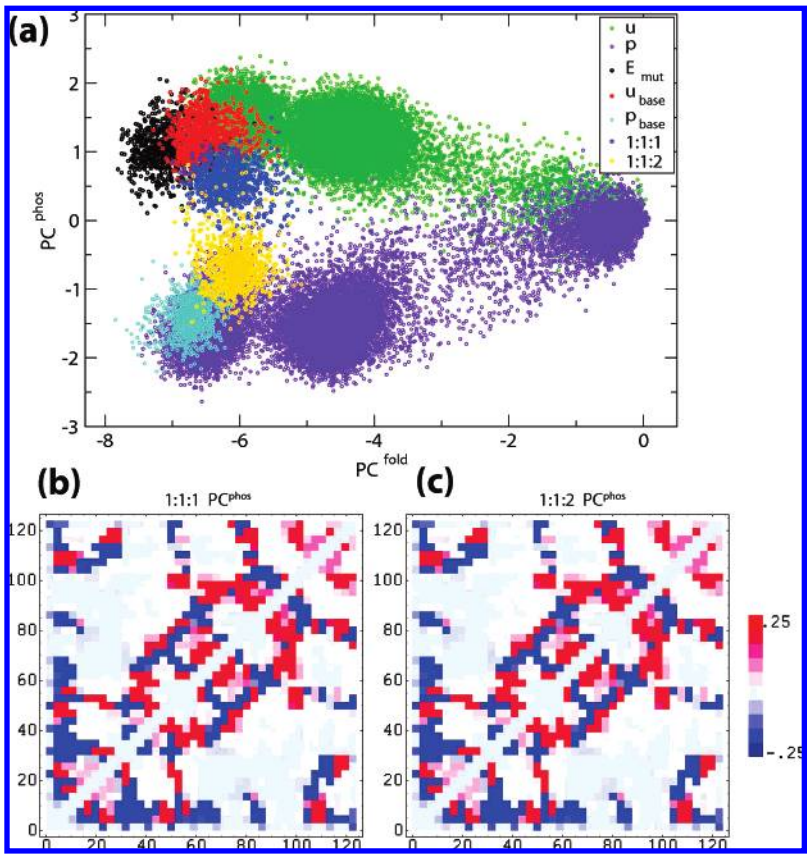

FIGURE 5: PCA of the contact maps for the conformations of NtrC obtained at $T = 0.75$ (the lower temperature facilitates sampling of the folded structures) with the native-structure-based Hamiltonian and also the phosphopredictive AMH. Also shown are the contact maps for phosphorylation principal component for the ensembles obtained with the phosphopredictive Hamiltonian with short:medium:long-range energy ratio of 1:1:1 and 1:1:2.

the phosphorylation event to act as a strict switch. For NtrC this barrier is several times larger and the only way for the unphosphorylated NtrC to reach the phosphorylated basin should be by means of more sophisticated pathways including local unfolding. In our view it is clear that protein cracking motions (*32*, *33*) are involved in the change.

PREDICTION OF STRUCTURAL CHANGES IN CYSTATIN WITH THE LINEAR RESPONSE METHOD

Small structural changes in protein conformations upon pertubation can be predicted by a linear response method, which relates the changes in residue−residue interactions of the unphosphorylated Hamiltonian to the phosphorylated Hamiltonian. Experiments for cystatin indicate only minor, and hence small, global conformational change upon phosphorylation (*14*). The main global changes of phosphorylation seen in the contact map in Figure 1 include different contacts of the helical region (residues 10−28 for helix 1) with the β-like structures (residues 34−38 for strand 1, 40−46 for strand 2, 50−60 for strand 3, 80−93 for strand 4 and 100−105 for strand 5). There are also local rearrangements of contacts in the β strand 4 and the preceding loop region (residues 68−80) including the phospho residue. These trends of structural changes were correctly captured by the PCA for the native-structure-based simulations (see PC^{fold} in Figure 2).

We applied the linear response method to estimate the structural changes on the contact map of cystatin upon phosphorylating the protein. The result of the prediction of the change of contact formation, $\langle \delta q_{i,j} \rangle$, is shown in Figure 7 as a contact map, which allows direct inspection of the residue−residue contact changes. The linear response method results are in excellent agreement with experiment. The global structural changes, i.e., the loss of contact formation between the helix and the β-like regions, were well captured. Further, the linear response method predicted the same local changes in the loop region around the phosphorylated residue as observed in experiments. Additionally, loss of loop contacts in residue region 65−75 were predicted. This region changes conformation and exhibits a 1.1 Å rms deviation of the phosphorylated native NMR structure from the unphosphorylated native NMR structure. It is clear that this linear response method developed to capture structural changes upon phosphorylation provides results consistent with experimental results.

PREDICTION OF THE PHOSPHORYLATED CONFORMATION WITH AN AMH-LIKE CONTACT POTENTIAL

It would be desirable to have a transferable Hamiltonian, that can predict the structure of any protein before and after phosphorylation from sequence information alone. Much progress toward de novo structure prediction has already been made by our group with techniques like those employed in ref (*20*) and by other groups with other styles of energy function (*34*, *35*). However, the proteins that change under phosphorylation, as we see, probably deviate from a strictly

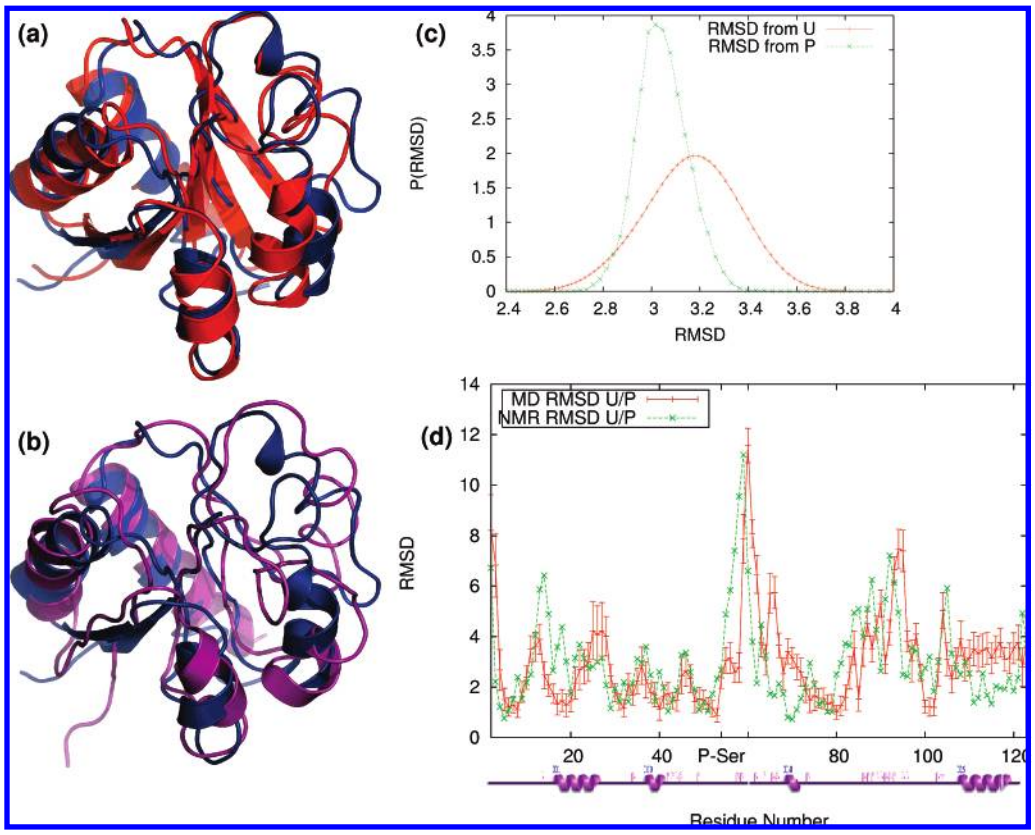

FIGURE 6: Overlay of a typical structure of NtrC (blue) obtained with the phosphopredictive Hamiltonian with the native NMR structure of the unphosphorylated form of NtrC (red) shown in (a) as well as the phosphorylated form of NtrC (purple). Probability distribution of rmsd from unphosphorylated (U) and phosphorylated (P) conformation shown in (c). In (d), rmsd as a function of residue index shown for the NMR structures as well as for the ensembles obtained from molecular dynamics simulations. The curves also display the error bars for the simulation results.

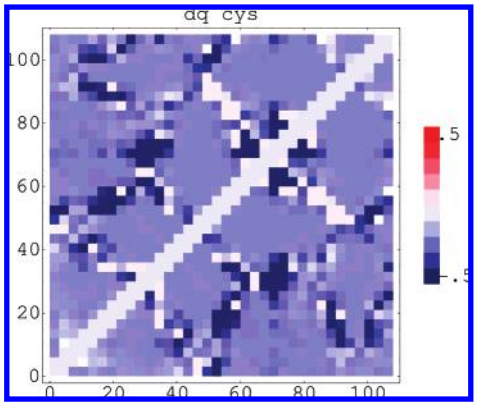

FIGURE 7: The linear response prediction of the changes of contact formation upon phosphorylation for cystatin.

funneled landscape. This makes the problem of complete de novo prediction more challenging than the usual. A much easier but still challenging computational problem would be to determine the structure of the phosphorylated test protein given only the structure of one form, say, the unphosphorylated conformation, or vice versa. Here we show how this can be done. To model how phosphorylation alters the tertiary structure of the protein conformation, we designed a predictive Hamiltonian $\mathscr{H}^*_p$, using short-range structural elements found in one form along with generic tertiary interactions. This Hamiltonian described in the method section is based on the de novo AMW prediction scheme. We call it the "phosphopredictive AMH''. The Hamiltonian $\mathscr{H}^*_p$ uses, as the sole input, the conformation of the unphosphorylated protein for only the short- and intermediate-range interactions. This assures a strong bias in the short and medium class for local secondary structure to form such elements as seen in the unphosphorylated protein. To model the effect of phosphorylation we have introduced a tunable 3-well long-range (in sequence space) residue−residue contact potential. This potential is modified to include interactions of the phospho residue. The strategy to model a phospho residue as a supercharged glutamic acid residue in the long-range potential can now be tested. We call the resulting energy function the "phosphopredictive'' Hamiltonian.

The first set of molecular dynamics simulations with the phosphopredictive Hamiltonian were performed with the original sequence of the unphosphorylated proteins, cystatin and NtrC. Since the input used is the contact map of the unphosphorylated protein, the predictive Hamiltonian mainly samples structures similar to those found in the folded basin of the unphosphorylated proteins when the long-range term is added as a perturbation term. Additionally, the energetic contributions of short-, medium- and long-range potentials were scaled to be equal in these simulations in keeping with estimates of the contributions of these parts of the interaction for funneled proteins. To check whether the sampled structures were similar to the structures found in the folded basins that would be obtained with the pure native-structure-based Hamiltonians, these snapshots were projected onto the first two principal components obtained with the native-structure-based Hamiltonians. The projections of the snap-

shots obtained with this Hamiltonian for NtrC are shown in red in Figure 5. Clearly the introduction of the long-range potential did not alter the ability to sample native unphosphorylated conformations. These projections serve as a baseline for the changes from results obtained with a pure native-structure-based Hamiltonian to those from a Hamiltonian with a heterogeneous contact potential.

Phosphorylation effects can be mimicked first by mutating the phospho residue simply to a glutamic acid. Thus a set of molecular dynamics simulations with the predictive Hamiltonian based on a pure were performed with precisely this modification in which the phospho residue was mutated to a glutamic acid. The snapshots for these simulations were projected onto the first two principal components and the results for NtrC were plotted in black in Figure 5. Clearly, the snapshots only slightly deviate from the snapshots of the folded state of the unphosphorylated protein. To test if using a nonadditive potential with water-mediated interactions will improve the quality of the prediction of the phosphorylated state, the same simulations were performed with the AMW potential (*36*). Contact maps of each snapshot obtained with the AMW were computed and projected onto the pricncipal components. The AMW ensemble projection had almost identical values of PC^{phos} and PC^{fold}, and hence contact formation, as did the ensemble obtained with the simple contact based phosphopredictive Hamiltonian for the same glutamic acid mutant. The rmsd's of heavy atoms of both the predicted ensembles from the NMR structure of the phosphorylated NtrC were similar. The AMW had on average 0.1 Å lower RMSDs from the NMR structure. Simulations with the AMW did show only minor improvement over the AMC in this case.

An important feature of our predictive Hamiltonian is the ability to "supercharge'' the phospho residue, that had been mutated into a glutamic acid. It is possible to assign different weights to the strength of interaction of the supercharged residue with other residues. Simulations have been performed for two different scalings of the strength of interaction, namely 1.4 and 2.0. The difference in results obtained with Hamiltonians of these two charge scales is subtle. We will explicitly show only the results for a charge of 1.4. The contact maps of the structures sampled with the supercharged phosphopredictive AMH were computed and projected onto the folding and phosphorylation principal components (see Figure 5, blue dots). The PC^{fold} values of the sampled conformations had similar PC^{fold} values to both the values of the unphosphorylated and phosphorylated ensembles. The more informative principal component, the phosphorylation principal component PC^{phos}, was shifted toward more negative values indicating enhanced formation of those contacts as seen in the phosphorylated ensemble rather than the unphosphorylated ensemble. To elucidate the predictive capability of the phosphopredictive Hamiltonian, the contact map corresponding to PC^{phos} was plotted (Figure 5). Defining four main helices in the native NMR structure of the phosphorylated form of NtrC (residues 15−27 correspond to helix 1, residues 36−42 to helix 2, residues 67−73 to helix 3 and residues 108−123 to helix 4), the contact map displays long-range contact changes for the phospho residue (residue 54) with the turn region before helix 1, and also between the regions of helix 3 and helix 4, that are similar to the changes in contact formation seen for the vanilla Hamiltonian. The contact changes between the phospho residue and the helix 2 region were not seen. Apart from those, the predictive Hamiltonian captured the long-range effects of the modified phospho residue in good agreement with the experimental determinations. To measure the quality of the structures sampled with the phosphopredictive AMH, the rmsd of the heavy atoms from their native NMR structure were computed. The average rms deviation from the NMR structure of the phosphorylated NtrC was about 2.7 Å with a standard deviation of 0.1 Å. Since the principal component analysis indicated a closer resemblance to the unphosphorylated ensemble rather than the phosphorylated ensemble, the rmsd of heavy atoms from the NMR structure of the unphosphorylated NtrC were also computed. The average rmsd was about 2.3 Å with a standard deviation of 0.1 Å. This result is not surprising due to the fact that the short- and medium-range structure is strongly biased toward the native structure of the unphosphorylated form. A more valid assesment of the quality of the predicted structure can be made by comparing the rmsd of the predicted ensemble from the respective ensembles, that would be obtained, when the NMR structures and sequences served as the sole input for the phosphopredictive Hamiltonian (see Figure 5, red dots for the unphosphorylated ensemble and cyan dots for the phosphorylated ensemble). We call these ensembles the baseline ensembles. Both baseline ensembles have similar projections on the principal component space compared with their respective ensembles obtained with the vanilla Hamiltonians. The predicted ensemble (blue) has an average of about 2.5 Å rmsd from both baseline ensembles, the phosphorylated and unphosphorylated ones.

The simulations, so far, were performed with short-, medium- and long-range contributions to the energy that are kept equal. This fact is motivated by the findings of Saven and Wolynes (*37*), who have estimated that in protein folding the contribution to the native energy arising from specific local interactions is comparable to those arising from specific tertiary interactions. It is therefore interesting to see whether different weights of the energetic contribution of the long-range interactions might improve the predictions. Several sets of simulations were performed with different total strength of interactions ranging from half the original strength up to twice as large. Most simulations did not show any better structures than what could be predicted using simulations of the glutamic acid mutant only. Only the results for simulations with twice the strength of the long-range interactions are therefore shown in Figure 5. These results display the most improvement for the prediction results. The contact maps were computed and projected onto the folding and phosphorylation principal components (see Figure 5, yellow dots). On a residue−residue contact level, this Hamiltonian best described the contact changes observed upon phosphorylation of NtrC. The scaled long-range interactions did perturb the local structure of the protein. An overlay of several predicted structures is shown in Figure 6a,b for visualization.

We also calculated for each molecular dynamics snapshot the rmsd of backbone atoms only from both NMR structures, the unphosphorylated and the phosphorylated NMR structures. Figure 6c shows the respective probability distributions. The probability distribution of the root-mean-square deviations of sampled structures from the NMR structure of the

phosphorylated conformation (Figure 6c, green curve) is seen to be shifted slightly toward lower values compared to the distribution of the deviation from the unphosphorylated NMR structure (Figure 6c, red curve). In other words, the structures obtained with the phosphopredictive AMH resemble more the phosphorylated form than they do the unphosphorylated form. In the range of observed values of rmsd, the phosphopredictive AMH clearly predicts structures which are more similar to the phosphorylated conformation than to the unphosphorylated conformation. These results support the findings of the principal component analysis that the phosphopredictive AMH indeed does predict conformations most similar to the global structure of the phosphorylated NMR structure.

Another way of seeing whether changes due to the phosphorylation are well predicted is to compare the motions at the individual residue level. We therefore computed two sets of rmsd at residue resolution. The first is the rmsd of the experimentally determined, unphosphorylated NMR structure from the experimentally determined, phosphorylated NMR structure. The second is the rmsd and its uncertainty for the predicted snapshots of the phosphorylated conformation from the structure of the average predicted structure of the unphosphorylated conformation. These comparisons allow us to directly compare the structural changes observed in simulation to the structural changes observed in experiment at the individual residue level. The results of these comparisons are plotted in Figure 6d, where the green curve represents the C^{α}-rmsd between the experimentally determined (unphosphorylated vs phosphorylated) structures and the red curve represents the difference for the corresponding conformations obtained in the molecular dynamics simulations. The trend of rmsd differences at the individual residue level observed in the experimentally determined structures and for the predicted structures shows the phosphopredictive AMH correctly captures the same structural changes at residue level that are observed in experiments. In particular the predicted regions of largest structural change correlate well with the most moved regions determined from experiment. We see the phosphopredictive AMH can provide a useful tool not only for theoreticans who wish to tease out the molecular forces responsible for phosphorylation induced conformational switching but also for those who only wish to identify at residue level and on a residue−residue contact level the effect of phosphorylation and thereby understand where key mutations offering phosphorylation induced changes can be made.

CONCLUSIONS

We have first discussed simulations with the native-structure-based Hamiltonians $\mathscr{H}_u$ and $\mathscr{H}_p$. While unphosphorylated and phosphorylated conformations both pre-exist on the landscape, these studies indicate the change of the landscape by post-translational modification is needed to allow the different structure ensembles to compete thermodynamically. To relate the landscapes of the two forms of a protein one can calculate the free energy differences using the cumulant expansion method (see Figures 2 and 3). The perturbation approach shows that phosphorylation changes the free energy profile by tilting the landscape to favor the phosphorylated basin. The calculations show the unphosphorylated protein has evolved so as not to adopt the phosphorylated conformation until the protein gets modified through phosphorylation even though superficially the rmsd between these conformations seems not to be very large. For a simply funneled completely minimally frustrated protein landscape idealized by our Hamiltonians $\mathscr{H}_u$ and $\mathscr{H}_p$, the unphosphorylated protein would rarely adopt the structure of the phosphorylated protein without post-translational modification. Partial unfolding mechanisms are likely required for these dramatic conformational switching events in NtrC.

Principal component analysis allows us to visualize the conformations of the ensembles of unphosphorylated and phosphorylated test proteins by projecting all changes onto the first two dominant components. In Figures 2 and 3, one contact based principal component PC^{phos} displayed mapping the major residue contacts that change upon phosphorylation. This contact map compares quite well to the contact map obtained from the linear response theory prediction of the changes (Figures 2 and 7).

Finally we used a structure prediction Hamiltonian, $\mathscr{H}^*_p$, to predict the final phosphorylated conformation for two systems. This algorithm successfully captures both the trends of conformational change of the unphosphorylated protein upon phosphorylation that are observed in experiments for the long-range contacts of the phospho residue and gives indeed the dominant structures. The phospho-predictive AMH provides a powerful tool to predict the major changes of structure upon phosphorylation given only information on the unphosphorylated conformation, or vice versa, pinpointing the major residue contact shifts. The Hamiltonian is general and captures the contact changes seen in small conformational changes as well as large conformational changes.

ACKNOWLEDGMENT

Additional computational support was provided in part by the NSF based Center for Theoretical Biological Physics.

REFERENCES

1. Hunter, T. (2000) Signaling−2000 and beyond, *Cell 100*, 113−127.
2. Johnson, L. N., and Barford, D. (1993) The effects of phosphorylation on the structure and function of proteins, *Annu. Rev. Biophys. Biomol. Struct. 22*, 199−232.
3. Steen, H., Jebanathirajah, J. A., Springer, M., and Kirschner, M. W. (2005) Stable isotope-free relative and absolute quantitation of protein phosphorylation stoichiometry by MS, *Proc. Natl. Acad. Sci. U.S.A. 102*, 3948−3953.
4. Radhakrishnan, I., PerezAlvarado, G. C., Parker, D., Dyson, H. J., Montminy, M. R., and Wright, P. E. (1997) Solution structure of the kix domain of cbp bound to the transactivation domain of creb: A model for activator:coactivator interactions, *Cell 91*, 741−752.
5. Buck, M., and Rosen, M. K. (2001) Flipping a switch, *Science 291*, 2329−2330.
6. Ramelot, T. A., and Nicholson, L. K. (2001) Phosphorylation-induced structural changes in the amyloid precursor protein cytoplasmic tail detected by NMR, *J. Mol. Biol. 307*, 871−884.
7. Johnson, L. N., and Lewis, R. J. (2001) Structural basis for control by phosphorylation, *Chem. Rev. 101*, 2209−2242.
8. Pufall, M., Lee, G. M., Nelson, M. L., K. H. S., Velyvis, A., Kay, L. E., McIntosh, L. P., and Graves, B. J. (2005) Variable control of ets-1 DNA binding by multiple phosphates in an unstructured region, *Science 309*, 142−145.
9. Park, K.-S., Mohapatra, D. P., Misonou, H., and Trimmer, J. S. (2006) Graded regulation of the kv2.1 potassium channel by variable phosphorylation, *Science 18*, 976−979.

10. Okamura, H., Aramburu, J., Garcia-Rodriguez, C., Viola, J. P. B., Raghavan, A., Tahiliani, M., Zhang, X., Qin, J., Hogan, P., and Rao, A. (2000) Concerted dephosphorylation of the transcription factor nfat1 induces a conformational switch that regulates transcriptional activity, *Mol. Cell 6*, 539−550.
11. Tholey, A., Pipkorn, R., Bossemeyer, D., Kinzel, V., and Reed, J. (2001) Influence of myristoylation, phosphorylation, and deamidation on the structural behavior of the N-terminus of the catalytic subunit of CAMP-dependent protein kinase, *Biochemistry 40*, 225−231.
12. Shen, T., Wong, C. F., and McCammon, J. A. (2001) Atomistic Brownian dynamics simulation of peptide phosphorylation, *J. Am. Chem. Soc. 123*, 9107−9111.
13. Groban, E. S., Narayanan, A., and Jacobson, M. P. (2006) Conformational changes in protein loops and helices induced by post-translational phosphorylation, *PLoS Comput. Biol. 2*, 238−250.
14. Dieckmann, T., Mitschang, L., Hofmann, M., Kos, J., Turk, V., Auerswald, E. A., Jaenicke, R., and Oschkinat, H. (1993) The structures of native phosphorylated chicken cystatin and of a recombinant unphosphorylated variant in solution, *J. Mol. Biol. 234*, 1048−1059.
15. Iakoucheva, L. M., Radivojac, P., Brown, C. J., O'Connor, T. R., Sikes, J. G., Obradovic, Z., and Dunker, A. K. (2004) The importance of intrinsic disorder for protein phosphorylation, *Nucleic Acids Res. 32*, 1037−1049.
16. Kern, D., Volkman, B. F., Luginbuhl, P., Nohaile, M. J., Kustu, S., and Wemmer, D. E. (1999) Structure of a transiently phosphorylated switch in bacterial signal transduction, *Nature 402*, 894−898.
17. Kern, D., Volkman, B. F., and Wemmer, D. E. (2001) A signaling protein 'in action'—structure and dynamics of a transiently phosphorylated switch, *Biophys. J. 80*, 13a.
18. Hardin, C., Eastwood, M. P., Luthey-Schulten, Z., and Wolynes, P. G. (2000) Associative memory hamiltonians for structure prediction without homology: Alpha-helical proteins, *Proc. Natl. Acad. Sci. U.S.A. 97*, 14235−14240.
19. Hardin, C., Eastwood, M. P., Prentiss, M., Luthey-Schulten, Z., and Wolynes, P. G. (2002) Folding funnels: The key to robust protein structure prediction, *J. Comput. Chem. 23*, 138−146.
20. Prentiss, M. C., Hardin, C., Eastwood, M. P., Zong, C. H., and Wolynes, P. G. (2006) Protein structure prediction: The next generation, *J. Chem. Theory Comput. 2*, 705−716.
21. Shen, T. Y., Zong, C. H., Hamelberg, D., McCammon, J. A., and Wolynes, P. G. (2005) The folding energy landscape and phosphorylation: modeling the conformational switch of the nfat regulatory domain, *FASEB J. 19*, 1389−1395.
22. Eastwood, M. P., Hardin, C., Luthey-Schulten, Z., and Wolynes, P. G. (2001) Evaluating protein structure-prediction schemes using energy landscape theory, *IBM J. Res. Dev. 45*, 475−497.
23. Hardin, C., Eastwood, M., Luthey-Schulten, Z., and Wolynes, P. G. (2003) Associative memory hamiltonians for structure prediction without homology:alpha/beta proteins, *Proc. Natl. Acad. Sci. U.S.A. 100*, 1679−1684.
24. Eastwood, M. P., Hardin, C., Luthey-Schulten, Z., and Wolynes, P. G. (2002) Statistical mechanical refinement of protein structure prediction schemes: Cumulant expansion approach, *J. Chem. Phys. 117*, 4602−4615.
25. Latzer, J., Eastwood, M. P., and Wolynes, P. G. (2006) Simulation studies of the fidelity of biomolecular structure ensemble recreation, *J. Chem. Phys. 125*, 214905-1−214905-12.
26. Ikeguchi, M., Ueno, J., Sato, M., and Kidera, A. (2005) Protein structural change upon ligand binding: Linear response theory, *Phys. Rev. Lett. 94*, 078102-1−078102-4.
27. Saito, N., Hashitsume, N., Toda, M., and Kubo, R. (2003) *Statistical Physics II: Nonequilibrium Statistical Mechanics*, 2nd ed., Springer, New York.
28. Jiang, Z. H. G., and McKnight, C. J. (2006) A phosphorylation-induced conformation change in dematin headpiece, *Structure 14*, 379−387.
29. Hupp, T. R., and Lane, D. P. (1995) Two distinct signaling pathways activate the latent DNA binding function of p53 in a casein kinase ii-independent manner, *J. Biol. Chem. 270*, 18165−18174.
30. Biswas, P., Zou, J. M., and Saven, J. G. (2005) Statistical theory for protein ensembles with designed energy landscapes, *J. Chem. Phys. 123*, 154908-1−154908-12.
31. Onuchic, J. N., LutheySchulten, Z., and Wolynes, P. G. (1997) Theory of protein folding: The energy landscape perspective, *Annu. Rev. Phys. Chem. 48*, 545−600.
32. Miyashita, O., Onuchic, J. N., and Wolynes, P. G. (2003) Nonlinear elasticity, proteinquakes, and the energy landscapes of functional transitions in proteins, *Proc. Natl. Acad. Sci. U.S.A. 100*, 12570−12575.
33. Ansari, A., Berendzen, J., Bowne, S. F., Frauenfelder, H., Iben, I. E. T., Sauke, T. B., Shyamsunder, E., and Young, R. D. (1985) Protein States and Proteinquakes, *PNAS 82*, 5000−5004.
34. Misura, K. M. S., Chivian, D., Rohl, C. A., Kim, D. E., and Baker, D. (2006) Physically realistic homology models built with rosetta can be more accurate than their templates, *Proc. Natl. Acad. Sci. U.S.A. 103*, 5361−5366.
35. Yang, J. S., Chen, W. W., Skolnick, J., and Shakhnovich, E. I. (2007) All-atom ab initio folding of a diverse set of proteins, *Structure 15*, 53−63.
36. Papoian, G. A., Ulander, J., Eastwood, M. P., Luthey-Schulten, Z., and Wolynes, P. G. (2004) Water in protein structure prediction, *Proc. Natl. Acad. Sci. U.S.A. 101*, 3352−3357.
37. Saven, J. G., and Wolynes, P. G. (1996) Local conformation signals and the statistical thermodynamics of collapsed helical proteins, *J. Mol. Biol. 257*, 199−216.

BI701350V