Ultraviolet Spectroscopy of Protein Backbone Transitions in Aqueous Solution: Combined QM and MM Simulations

Jun Jiang,[†] Darius Abramavicius,[†] Benjamin M. Bulheller,[‡] Jonathan D. Hirst,[‡] and Shaul Mukamel*,[†]

Chemistry Department, University of California Irvine, Irvine, California, and School of Chemistry, University of Nottingham, University Park Nottingham NG7 2RD, United Kingdom

Received: March 4, 2010; Revised Manuscript Received: May 12, 2010

A generalized approach combining quantum mechanics (QM) and molecular mechanics (MM) calculations is developed to simulate the $n \rightarrow \pi^*$ and $\pi \rightarrow \pi^*$ backbone transitions of proteins in aqueous solution. These transitions, which occur in the ultraviolet (UV) at 180−220 nm, provide a sensitive probe for secondary structures. The excitation Hamiltonian is constructed using high-level electronic structure calculations of *N*-methylacetamide (NMA). Its electrostatic fluctuations are modeled using a new algorithm, EHEF, which combines a molecular dynamics (MD) trajectory obtained with a MM forcefield and electronic structures of sampled MD snapshots calculated by QM. The lineshapes and excitation splittings induced by the electrostatic environment in the experimental UV linear absorption (LA) and circular dichroism (CD) spectra of several proteins in aqueous solution are reproduced by our calculations. The distinct CD features of α-helix and β-sheet protein structures are observed in the simulations and can be assigned to different backbone geometries. The fine structure of the UV spectra is accurately characterized and enables us to identify signatures of secondary structures.

Introduction

The function of proteins normally depends crucially on their secondary structure and dynamic fluctuations.[1] Optical spectroscopy provides a direct probe of the conformations of biological systems and the coupling between biomolecules and their surroundings.[2,3] The investigation of these systems requires the development of simulation tools that adequately represent the fluctuations of the molecular environment. Protein motions in aqueous solution lead to fluctuations of the electrostatic environment. These, in turn, induce changes in the intra- and intermolecular interactions and thereby change the local Hamiltonian. Hamiltonian fluctuations shift and broaden the spectra, and a proper description is crucial for the prediction of spectral fine structure and lineshapes. To simulate the fluctuation effects, one might need to consider thousands of snapshots to reflect the conformation diversity.[4] However, repeated quantum mechanics (QM) calculations are prohibitively expensive. QM approaches can accurately describe small molecules, where small typically means <100 atoms. Molecular mechanics (MM) methods can describe large complexes but neglect potentially important quantum mechanical effects. The combined approach of QM and MM calculations has become widely used for simulating large biological molecules.[5,6] A typical combined simulation generates geometric snapshots along the molecular dynamics (MD) trajectory based on a MM forcefield and applies QM methods to describe the electronic structure for each snapshot. For example, accurate QM theory can be used to describe the active site of a protein, whereas the contribution of the rest of the system is treated more approximately.

Significant progress has been made in the simulation of vibrational infrared spectra of proteins by building a map to represent the local Hamiltonian as a function of geometric parameters or electric fields at some reference points.[4,7,8] Such "map function" methods reproduce the electrostatic environment from many MD snapshots with reasonable calculation cost. In such approaches, the physical relationships between the Hamiltonian and geometrical structures are replaced by some fitting functions. The simulation results are sensitive to the number of reference points, the way they are chosen, and the type of functions used. Because of the lack of simple physical guidelines, complicated numerical analysis and expensive test calculations must be performed to establish a special map for a specific molecules. It is hard to construct a generalized map model that is transferable between different systems.

Linear absorption (LA) and circular dichroism (CD) spectroscopy of proteins in the ultraviolet (UV) region 180−220 nm are commonly used for secondary structure characterization.[9] Two important types of secondary structure elements are α-helix and β-sheet. Specific regions in the spectra reflect the electronic excitations. In CD spectra, α-helical proteins show a strong positive peak at 190 nm (52 000 cm^{-1}) and a negative doublet at 208 and 222 nm (48 000 and 45 500 cm^{-1}). Sheet-containing proteins are less ordered, and the CD spectra vary a little more. Their common features are a negative amplitude at 180 nm (55 000 cm^{-1}), a positive band at ∼195 nm (51 000 cm^{-1}), and usually a negative peak at ∼215 nm (44 500 cm^{-1}).[10] The simulation of UV spectra by the matrix method has been shown to be very successful by several groups.[11−14] On the basis of either empirical parametrization or ab initio parameter sets, the DichroCalc package[15,16] has given good agreement between simulated and experimental CD spectra.

In previous simulations, fluctuations were added phenomenologically by convoluting the spectra with a broadening factor.[13,17] Here we develop a generalized approach combining

* To whom correspondence should be addressed. E-mail: smukamel@uci.edu.
[†] University of California Irvine.
[‡] University of Nottingham

10.1021/jp101980a
Published on Web 05/26/2010

QM and MM methods for calculating these spectra. We focus on the simulation of the $n \rightarrow \pi^*$ and $\pi \rightarrow \pi^*$ (denoted as $n\pi^*$ and $\pi\pi^*$ below) bands of the protein backbone in the 180−220 nm wavelength region. We start with high-level electronic structure calculations on *N*-methylacetamide (NMA), which is a model system for the peptide bond. Two electronic excitations in NMA have been considered: the $n\pi^*$ and $\pi\pi^*$ transitions. Using this as a basis, we constructed the exciton Hamiltonian. For the proteins, MD simulations in water were performed to create a large number of geometric snapshots. A number of these snapshots were selected as representative structures. QM calculations were carried out to compute the ground-state electronic density of the protein. To combine the QM and MM simulation results, we have developed an efficient algorithm called exciton Hamiltonian with electrostatic fluctuations (EHEF). EHEF performs charge population analysis for the MD samples. Charges contributed by localized atomic orbitals are treated as atomic partial charges, and charges arising from delocalized atomic orbitals are treated with a set of grid point charges fitted from the electrostatic potential. A set of standard atom−atom charges are generated in the "internal coordinate frame". For a given conformation, charge distributions were deduced from the standard atom−atom charges by updating atom−atom vectors of the corresponding MD geometric structure. Using the full charge distributions, we calculate the interactions between the chromophore and environment. We, thus, avoid expensive repeated QM calculations and obtain the fluctuating Hamiltonian at the QM level for all MD snapshots. The present algorithm is based on physical considerations that require no empirical parameters and can be transferred directly to other systems. Using this algorithm, we have studied the UV LA and CD spectra for several typical proteins in aqueous solution: hemoglobin, leptin, tropomyosin, lentil lectin, monellin, and FtsZ. The very distinctive features of UV spectra, which depend on the secondary structure, are reproduced in good agreement with experiment.

Theoretical Methods

Exciton Model for the $n\pi^*$ and $\pi\pi^*$ Transitions. Protein backbone electronic transitions can be described by the Frenkel exciton model[18,19] in the Heitler−London approximation.

$$\hat{H} = \sum_{ma} \varepsilon_{ma} \hat{B}^{\dagger}_{ma} \hat{B}_{ma} + \sum_{ma,nb}^{m \neq n} J_{ma,nb} \hat{B}^{\dagger}_{ma} \hat{B}_{nb} \qquad (1)$$

where *ma* is the *a* electronic transition on the peptide unit, *m* (in our case, $a = 1$ for $n\pi^*$ and $a = 2$ for $\pi\pi^*$). $\hat{B}^{\dagger}_{ma}$ is the creation operator that promotes the *m* peptide unit into the excited state, *a*, and $\hat{B}_{ma}$ is the corresponding annihilation operator denoting the ground state $|0\rangle$. The commutation relations of these operators are $[\hat{B}_{ma}, \hat{B}^{\dagger}_{nb}] = \delta_{mn}(1 - 2\hat{B}^{\dagger}_{mb}\hat{B}_{ma})$.[20] The ground-state energy is $\langle 0|\hat{H}|0\rangle = 0$. In the single-exciton manifold, the *m*th singly excited state energy can be calculated as $\langle 0|\hat{B}_{ma}\hat{H}^{(e)}\hat{B}^{\dagger}_{ma}|0\rangle = \varepsilon_{ma}$, and the resonant coupling between singly excited states *m* and *n* is given by $\langle 0|\hat{B}_{ma}\hat{H}^{(e)}\hat{B}^{\dagger}_{nb}|0\rangle = J_{ma,nb}$.

By diagonalizing the Frenkel Hamiltonian matrix, we obtain the excitation energies and transition moments, which are then used to simulate the UV spectra. This model enables us to compute single excitations using QM calculations for each isolated chromophore of the entire protein. The resonant coupling between transition densities of two chromophores *m* and *n* is

$$J_{ma,nb} = \frac{1}{4\pi\varepsilon\varepsilon_0} \int\int d\mathbf{r_m}\, d\mathbf{r_n} \frac{\rho^{eg}_{ma}(\mathbf{r_m})\rho^{ge}_{nb}(\mathbf{r_n})}{|\mathbf{r_m} - \mathbf{r_n}|} \qquad (2)$$

where $\mathbf{r}$ is the spatial coordinate. $\rho^{eg}_{ma}(\mathbf{r_m})$ and $\rho^{ge}_{ma}(\mathbf{r_m})$ are the transition charge densities. All of the excitation energies and charge densities can be obtained from the QM calculations of the isolated chromophore. Because there are two dominant transitions in the far-UV region for proteins, we here consider two transitions $n\pi^*$ and $\pi\pi^*$ in each amide chromophore site (i.e., peptide bond).

To compute intermolecular couplings via eq 2, we need to calculate the permanent and transition charge densities of each molecule. We selected NMA as a model for an isolated peptide unit. The electronic excited states of NMA were taken from calculations,[15] using the complete-active space self-consistent-field method within a self-consistent reaction field (CASSCF/SCRF) and multiconfigurational second-order perturbation theory (CASPT2), as implemented in MOLCAS.[21] Monopoles for a given state were determined by fitting their electrostatic potential to reproduce the ab initio electrostatic potential for that state.[15] An ab-initio-based parameter set was extracted[15] to represent the transition energy and the permanent and transition charge densities of the isolated peptide unit.

Electrostatic Fluctuations. As recognized in a previous work of Kurapkat et al., interactions of a chromophore with local electrostatic fields can lead to considerable energy shifts of its transitions.[22] The transition energy is affected by the fluctuating electrostatic potential coming from the rest of the protein and the surrounding solvent. These effects will be incorporated below.

To set the stage, we first survey some commonly used methods. In the dipole approximation, the state energy, ε, can be expressed as[23]

$$\varepsilon = \varepsilon_0 + \mu \cdot \mathbf{F} \qquad (3)$$

where ε_0 represents the state energy of the isolated molecule, μ is the electric dipole moment, and $\mathbf{F}$ is the electric field induced by the surroundings. The transition energy, ε^F_{ma}, including electrostatic environmental fluctuations, is then computed to be

$$\begin{aligned} \varepsilon^F_{ma} &= (\varepsilon^m_0 - \varepsilon^g_0) - (\mu^{ee}_{m,a} - \mu^{gg}_m) \cdot \mathbf{F} \\ &= \varepsilon_{ma} - (\mu^{ee}_{m,a} - \mu^{gg}_m) \cdot \mathbf{F} \end{aligned} \qquad (4)$$

where $|g\rangle$ denotes the ground state and μ^{ee}_{ma} and μ^{gg}_m are the permanent dipoles of the excited states and ground state, respectively. Nevertheless, the above formula is not very accurate for extended systems. The main problem is that the dipole moment and electric field are not evenly distributed in space, so that a single $(\mu^{ee}_{m,a} - \mu^{gg}_m)\cdot\mathbf{F}$ factor cannot account for the environment fluctuation corrections to transition energies. To account for the spatial distribution of μ^{ee}_{ma}, μ^{gg}_m, and $\mathbf{F}$, one can use a set of reference points in the peptide and build a map to represent the excitation energy as a function of the electric field at those reference points. To represent the spatial distributions better, the gradient or higher order derivatives of the electric field may be used as variables. A simple map can be expressed as

$$\varepsilon_m^F = \varepsilon_m + \sum_i \alpha_i \mathbf{F}_i + \sum_i \beta_i \frac{d\mathbf{F_i}}{dr} + ... \quad (5)$$

$\mathbf{F}_i$ is the electric field at the ith reference point and α_i and β_i are empirical parameters obtained by fitting to experimental data. The local geometric changes of the excited chromophore are usually the main factors that affect the local Hamiltonian. Because obtaining the transition dipole moment requires expensive QM calculations, one can consider another type of map that parametrizes the excitation energy with geometric variables at reference points

$$\varepsilon_m^F = \varepsilon_m + \sum_i \alpha'_i \mathbf{K_i} + \sum_i \beta'_i \mathbf{K_i^2} + ... \quad (6)$$

where K_i stands for different geometric variables, such as atomic bond lengths, bond angles, dihedral angles, and so on. α'_i and β'_i are fitted parameters.

Such map methods avoid the expensive QM calculations for the excited chromophores under the influence of the environment. A major limitation is that the parameters can only be obtained by fitting theoretical results with experiments or high-level QM calculations. The simulations depend on the number and choice of reference points and the functions used to describe them. It is not possible to develop a universal map that is transferable between different systems.

Here we develop an alternative approach to calculate the full-space corrections of excitation energies due to electrostatic fluctuations. Instead of using the dipole moment, we compute the product of the transition charge density and electric field. By integrating that product over space, we calculate the interactions between the excited states and environment directly. The excitation energy corrected by the environment electrostatic potential and intermolecular interactions is then expressed as

$$\varepsilon_{ma}^F = \varepsilon_{ma} - \int d[\rho_{ma}^{ee}(\mathbf{r}) - \rho_m^{gg}(\mathbf{r})] \cdot \mathbf{r} \cdot \mathbf{F}(\mathbf{r}) \quad (7)$$

where ε_{ma} is the excitation energy of the ath transition of the chromophore, m. $\rho_{ma}^{ee}(\mathbf{r_m})$ and $\rho_m^{gg}(\mathbf{r_m})$ represent the molecular charge density of the excited and ground state of the chromophore, respectively. The electric field $\mathbf{F}(\mathbf{r})$ is computed as the gradient of the Coulomb potential induced by the ground-state charge density of the surrounding environment on the excited chromophore. The fluctuating excitation energy is thus given as

$$\varepsilon_{ma}^F = \varepsilon_{ma} + \sum_l \frac{1}{4\pi\varepsilon\varepsilon_0} \int\int d\mathbf{r_m}\, d\mathbf{r_l} \frac{[\rho_{ma}^{ee}(\mathbf{r_m}) - \rho_m^{gg}(\mathbf{r_m})]\rho_l^{gg}(\mathbf{r_l})}{|\mathbf{r_m} - \mathbf{r_l}|} \quad (8)$$

where l runs over the molecular sites surrounding the excited chromophore, m, and $\rho_l^{gg}(\mathbf{r_l})$ represents the charge density of the ground state of molecular site, l.

Simulation of the Full Ground-State Charge Distribution. The electronic structures of amino acid side chains in proteins and the surrounding water molecules were computed in the gas phase with density functional theory (DFT) implemented in the Gaussian03 package[24] at the B3LYP/6-311++G** level. The fragments considered were, for example, methane (representing the side chain of alanine), indole (representing the side chain of tryptophan), and so on and an individual water molecule. The full charge distribution is calculated from the DFT densities. In QM approaches, one can calculate charge densities by coarse-graining the electronic wave function in space. Krueger et al.[25] have employed a grid technique to compute the intermolecular couplings. In their transition density cube (TDC) method, 3-D space is divided into many small volume elements. Couplings were calculated directly from the Coulomb interactions between charges of the cubes in each molecule. This method gives high accuracy because it is based on full QM calculations. However, it is very expensive because it requires a large number of cubes to maintain the precision (normally ~500 000 cubes for a system with ~50 atoms). Madjet et al.[26] have developed a more efficient way for taking the charge distribution into account. In their TrEsp (transition charge from electrostatic potential) code, electrostatic potentials on sample points are computed at the QM level. Partial charges are then assigned to atomic positions by fitting to potentials. TrEsp has been successful in the study of many biological systems. However, charges distributed only at the atomic positions cannot represent the electronic cloud over the space and the corresponding electronic properties when the molecular orbitals are delocalized.

Our algorithm combines the advantages of the above methods to obtain affordable and accurate full charge distributions. On the basis of DFT calculations, we decomposed the Kohn−Sham orbitals into atomic orbitals and divided them into two groups: localized and delocalized. For the localized atomic orbitals, we adopted the TrEsp procedure; that is, we compute the electrostatic potential and fit it to atomic partial charges. We further computed charge distributions induced by the delocalized atomic orbitals, which vary in space much more slowly than the localized ones. We then introduce a set of regular grids and assign fictitious charges on the grid points. Each grid point is divided into a large number of small cubes, in a similar way to (and of a similar size to) cubes employed in the TDC method. Sample points around each grid point are taken, and the electrostatic potential induced by the charges of the small cubes inside that grid is calculated. The fictitious charge for each grid point is obtained by fitting the sample electrostatic potential. In the end, the atomic partial and fictitious grid charges are used to compute the Coulomb interactions and transition dipoles. Most localized charges are included as atomic partial charges, and the spatial distribution of those delocalized charges varies slowly with the coordinates so that the space resolution requirement is very loose. Distribution of delocalized charges can be described by a limited number of grid points (normally ~10 000 cubes for a system with ~50 atoms), which is much smaller than the number of cubes used in the TDC methods. Because the size of the grid points is much smaller than the interatomic distance, the resultant electrostatic interactions are as accurate as the TDC. This algorithm offers a good balance between accuracy and cost.

Atom−Atom Charges for Individual MD Snapshots. For molecules with a rigid geometry, one can define the charge distributions in a molecular frame and reuse them for any MD snapshots. However, proteins are not rigid, and the interatomic distances and angles do vary during the MD simulations. Nevertheless, the conformation dynamics only lead to slight changes in atomic bond lengths and angles. For a pair of atoms, one can expect that the distribution of their electronic charges changes only slightly with their atomic positions. We, therefore, define an "internal coordinate frame" for each pair of atoms

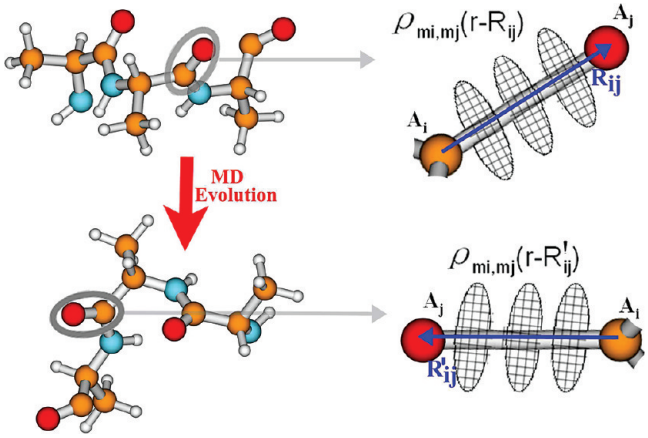

Figure 1. A_i and A_j have been selected to show the description of atom−atom charges based on the atom−atom vectors $\mathbf{R}_{ij}$ and $\mathbf{R}'_{ij}$, whose atom−atom charges are described as $\rho^{12}_{A_i,A_j}(\mathbf{r}-\mathbf{R}_{ij})$ and $\rho^{12}_{A_i,A_j}(\mathbf{r}-\mathbf{R}'_{ij})$, respectively.

and describe their charge distributions as atom−atom charges with respect to the atom−atom vector (vector between the atoms).

Standard QM methods compute the electronic charge density by integrating atomic orbitals over space

$$\rho_m^{12}(\mathbf{r}) = V_\delta \int_r^{r+\delta} \int_s \sum_\eta^{\text{occ}} \sum_{i,j} c_i^{1,\eta} c_j^{2,\eta} \psi_i \psi_j \, \mathrm{ds}\, \mathrm{d}\mathbf{r}' \qquad (9)$$

where $\rho_m^{12}(\mathbf{r})$ represents the transition charge density between state $|1\rangle$ and $|2\rangle$ of molecule, m. In this study, $|1\rangle$ and $|2\rangle$ are both limited to the ground state of amino acid side chains or water molecules to calculate their permanent charge densities. In principle, eq 9 is a general expression for both permanent and transition charge densities. V_δ is the volume element. $c_i^{1,\eta}$ and $c_j^{2,\eta}$ are the orbital coefficients of states $|1\rangle$ and $|2\rangle$, respectively, in which η runs over all occupied molecular orbitals. $\psi_i(\mathbf{r}-\mathbf{r}_A)$ is one of the basis functions of atom A_i. The charge density can, therefore, be decomposed in the atomic site representation

$$\rho_m^{12}(\mathbf{r}) = \sum_{i,j} V_\delta \int_r^{r+\delta} \int_s \sum_\eta^{\text{occ}} c_i^{1,\eta} c_j^{2,\eta} \psi_i(\mathbf{r}-\mathbf{r}_{A_i}) \psi_j(\mathbf{r}-\mathbf{r}_{A_j}) \, \mathrm{ds}\, \mathrm{d}\mathbf{r}' = \sum_{i,j} \rho^{12}_{A_i,A_j}(\mathbf{r}-\mathbf{R}_{ij}) \qquad (10)$$

in which $\mathbf{R}_{ij} = \mathbf{r}_{A_i} - \mathbf{r}_{A_j}$ is the atom−atom vector and A_i,A_j is a pair of atoms. The sum runs over all atomic positions in the molecule. $\rho^{12}_{A_i,A_j}(\mathbf{r}-\mathbf{R}_{ij})$ is the density arising from the atom−atom charges. A fragment of a protein is shown in Figure 1 to illustrate how we compute atom−atom charges. For every A_i,A_j pair, we define the atom−atom vector $\mathbf{R}_{ij}$. We also define a set of (usually ten) planar disks perpendicular to $\mathbf{R}_{ij}$, whose centers lie on $\mathbf{R}_{ij}$, located in the range $\pm 1.5|\mathbf{R}_{ij}|$. Each disk is divided into (typically 20 to 50) small grid points. The atom−atom charge for each pair of atoms A_i and A_j is computed to be $\int \rho^{12}_{A_i,A_j}(\mathbf{r}-\mathbf{R}_{ij})\, \mathrm{d}\mathbf{r}$. This was found to be stable during the MD simulations. The spatial distributions of atom−atom charges are described by their relative position with respect to the atom−atom vector, $\mathbf{R}_{ij}$.

We selected ideal structures or several sampled MD snapshots as representative structures. From the QM calculations, we computed the distributions of the atom−atom charges for the representative geometry. The standard atom−atom charges are generated for a representative atom−atom vector, $\mathbf{R}_{ij}$. When the geometry is varied and the atom−atom vector changes to $\mathbf{R}'_{ij}$, a new set of atom−atom charges, $\rho^{12}_{A_i,A_j}(\mathbf{r}-\mathbf{R}'_{ij})$, is mapped out from the standard ones by reproducing the position relative to $\mathbf{R}'_{ij}$. As long as the chemical structure does not vary strongly, the newly mapped atom−atom charges can accurately reflect the electronic properties of each MD snapshot at the QM level. For each MD snapshot, we calculated the full atom−atom charge distributions from the standard atom−atom charges. With the full atom−atom charge distribution, we calculated the electrostatic potential over all space, thus generating the local Hamiltonian for each MD snapshot. QM accuracy is retained, with very few QM calculations.

Computational Details

MD simulations of several proteins were performed in water with the CHARMM22 force field[27] and the TIP3P water model[28] in the software package NAMD.[29] We considered a dilute solution. Each residue feels the electrostatic potential from the surrounding water and other residues of the same protein. Simulations were conducted in the NPT ensemble, and we employed cubic periodic boundary conditions. The particle mesh Ewald sum method was used to treat the long-range electrostatics. A nonbonded cutoff radius of 12 Å was used. All MD simulations started from a 5000 step minimization and 600 ps heating from 0 K to room temperature, 310 K. The MD simulation time step was 1 fs. After 2 ns of equilibration, we simulated 16 ns dynamics at 1 atm pressure and 310 K. Structures were recorded every 400 fs. An ensemble of MD snapshots was used to compute the local excitation Hamiltonian and the UV spectra. The effect of the electrostatic potential generated by water, the peptide groups, and the amino acid side chains was investigated. The protein structures are stable during the MD simulations, with a root-mean-square deviation (rmsd) of backbone atoms from the initial structure of 0.7 to 1.5 Å and 0.7 to 2.4 Å for α-helix and β-sheet proteins, respectively.

QM calculations were performed using Gaussian03[24] and MOLCAS7.[21] Our EHEF algorithm was used to calculate the full atom−atom charge distribution from the QM calculations. The use of the exciton matrix method implemented in the DichroCalc program has been very successful in reproducing protein UV spectra.[13,14] Parameters for the transition energies of the isolated peptide units, the resonant couplings and electric and magnetic transition dipole moments are extracted from the DichroCalc package. By combining our SPECTRON code[20] with DichroCalc, we constructed the effective electronic excitation Hamiltonian and calculated the UV spectra. The simulated spectra reported here are based on 2000 MD snapshots, and they are compared with spectra computed with DichroCalc for a single conformation.[13]

UV Spectra of Proteins

Transition-Energy Fluctuations. We first examine how the fluctuations of the molecular environment affect the transition energy of a individual peptide group (ε^F_{ma} in eq 8). For the helical protein hemoglobin (RSCB code: 1hda) and sheet protein lentil lectin (RSCB code: 1les), the distributions of excitation energy ε^F_m of many peptide groups at 310 K relative to that of an isolated NMA have been depicted in Figure 2A,B. There are no clear

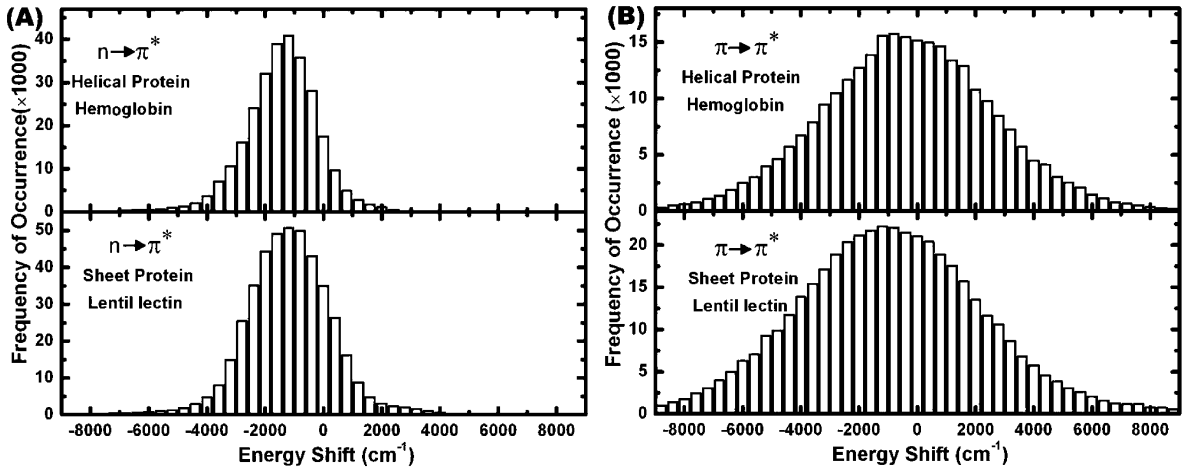

Figure 2. Distribution of transition energy shifts due to the electrostatic environment from 310 K simulations: (A) $n\pi^*$ (~45 454 cm^{-1}) transitions and (B) $\pi\pi^*$ (~52 631 cm^{-1}) transitions.

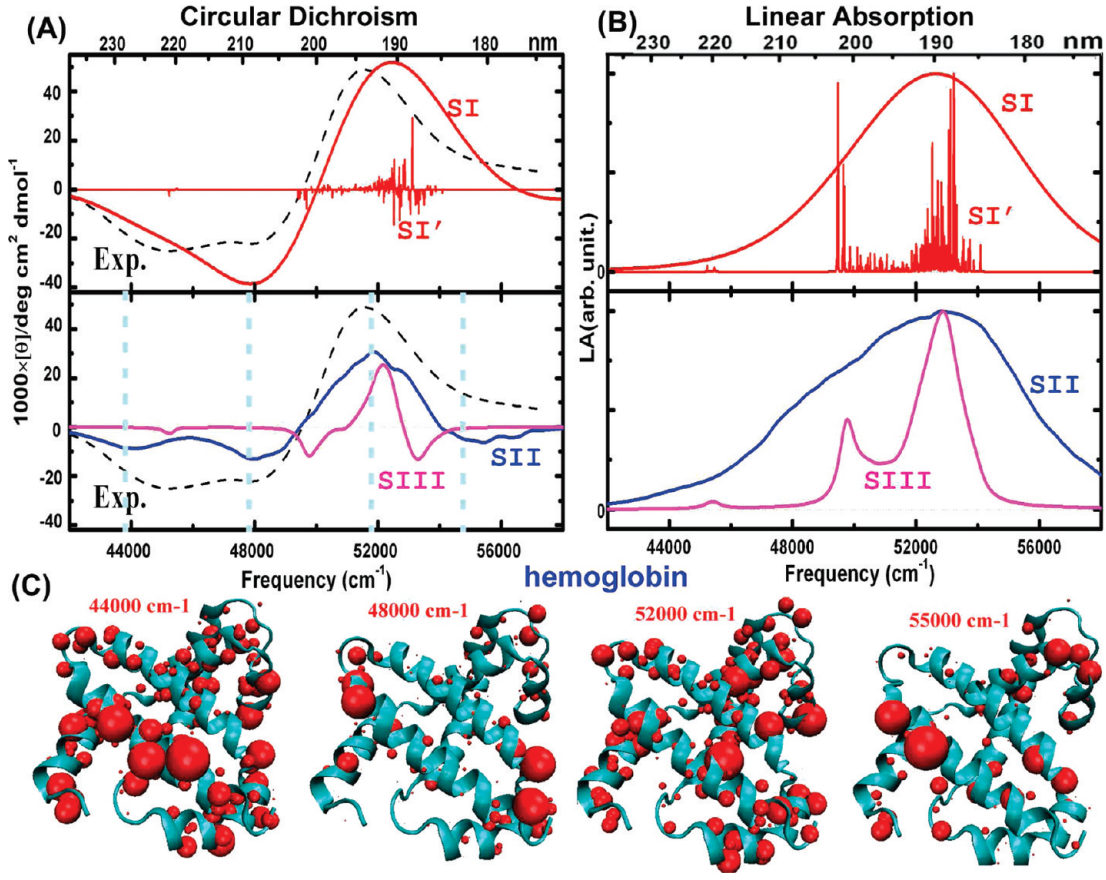

Figure 3. α-Helical protein hemoglobin (PDB code: 1hda) together with X-ray crystal structures. SI′ represents the simulated line spectra based on a single conformation (scaled by a factor 1/1000), and SI indicates SI′ convoluted with a Gaussian envelope. SII are simulated spectra based on 2000 MD snapshots that consider the electrostatic potential from all surroundings, and SIII takes account of only peptide groups (scaled by a factor 1/15). The experimental CD spectrum[31] is shown as dashed black lines. (C) Transition populations corresponding to the four CD peaks. The volume of the balls represents the amplitude of the electronic transition.

differences between the transition-energy fluctuations of helical and sheet proteins. Both the $n\pi^*$ and $\pi\pi^*$ transitions show distinct asymmetric distributions. The $n\pi^*$ and $\pi\pi^*$ transitions have red shifts in their excitation energies of ~1200 (0.15 eV) and ~1000 cm^{-1} (0.12 eV), respectively, which are consistent with a previous combined QM and MM study on the NMA molecule.[30] Our method is based on the full charge distributions, which are more accurate than the atomic charges obtained from Mulliken population analysis previously considered.[30] The distribution of $n\pi^*$ excitation energy shifts is similar to that computed from NMA in water,[30] but that of the $\pi\pi^*$ transition is much broader in our simulations. Different peptide groups are affected differently by the electrostatic potentials, which can vary from protein to protein. This is one reason why the bands in CD spectra of proteins can vary in their precise location.[13,14,31−33] It corresponds to the shifts of the $\pi\pi^*$ excitation energy of −8000 to 6000 cm^{-1} with respect to the excitation energy of the NMA molecule at 52631 cm^{-1}.

α-Helical Proteins: Hemoglobin. Hemoglobin is a typical α-helical protein. The structure is shown in Figure 3. Its X-ray

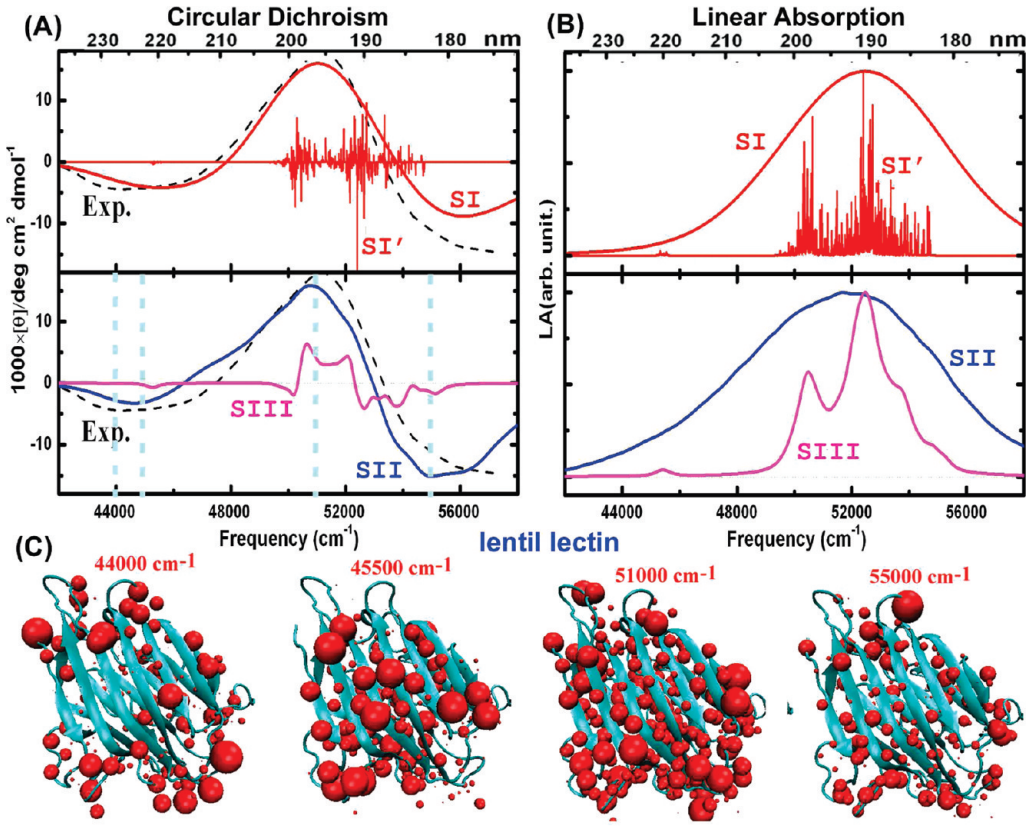

Figure 4. Same as for Figure 3, but for sheet-containing protein lentil lectin (PDB code: 1les).

crystal structure reported in the RSCB protein data bank (1hda) was taken as the starting geometry. MD simulations were carried out on the tetrameric hemoglobin, neglecting the heme groups. The UV spectra were calculated using a single chain geometry extracted from the MD trajectories. First, CD and LA line spectra calculated based on the single X-ray structure are plotted (labeled SI′) in Figure 3A,B. Spectra convoluted with Gaussian line shape with a full width at half-maximum height (fwhm) value of 12.5 nm[13,14] for the single conformation are plotted as well (labeled SI). The SI CD spectrum reproduces the experimental peaks[31] at 48 000 and 52 000 cm^{-1} but underestimates the intensity of the peak at ~44 000 cm^{-1}. The relationship between the CD line spectra and corresponding convoluted spectra is not very straightforward. For instance, a positive (negative) peak in the line spectrum can be shifted or even canceled in the convoluted spectrum by the negative (positive) contributions of neighboring transitions. The negative CD peak at 48 000 cm^{-1} in the SI in Figure 3A mainly arises from the negative CD signals at ~45 000 and 50 000 cm^{-1}. Moreover, in the LA spectra in Figure 3B, the transitions at 50 000 cm^{-1} evident in the line spectrum SI′ are buried in SI by the convoluted signals from 52 000 to 54 000 cm^{-1}, which are much stronger and denser in the frequency region.

The full CD spectrum obtained by using our algorithm for 2000 MD snapshots is shown as the curve SII in Figure 3A. SII provides a better resolution than SI of the experimentally observed double minimum. The experimental line shape is well described by this combined QM and MM simulation. The three main CD peaks at 44 000, 48 000, and 52 000 cm^{-1} are reproduced by SII. The origin of additional CD peaks in SII compared with SI is that peptide groups are affected by different electrostatic potentials. The SII LA spectrum shown in Figure 3B reproduces the bandshapes of the convoluted LA spectrum of the single conformation (SI LA).

To examine how the environment influences the electronic transitions, we have calculated spectra taking into account only the local fluctuations of peptide groups and neglecting the electrostatic potential induced by the amino acid side chains and surrounding water molecules. The simulated spectra are labeled SIII in Figure 3A,B. The SIII bandwidths are much narrower than SII. Protein backbone fluctuations account for only ~6 nm fwhm, whereas the amino acid side-chain fluctuations contribute ~10−12 nm. No empirical parameters were needed in generating SII to describe the inhomogeneous broadening. For the $\pi\pi^*$ transitions at ~52 000 cm^{-1}, the fwhm obtained from SII CD spectra is 10.4 nm, agreeing well with the fwhw of 12.0 nm found in the experimental spectra. In both experiment and simulation, the band of $n\pi^*$ transitions at 44 000 cm^{-1} overlaps with the band at 48 000 cm^{-1} (induced by exciton splitting of the $\pi\pi^*$ transitions) so that we have to extract their fwhm by fitting with Gaussian lineshapes separately. The fwhm of 44 000 and 48 000 cm^{-1} bands in the simulated spectra are found to be 14.5 and 10.2 nm, respectively, which are close to the values of 19.4 and 9.5 nm obtained from corresponding experimental bands. Additive simulations (not shown) found that the electrostatic potential induced by water only weakly affects the UV spectra. Water has only an indirect influence on the electronic transitions by affecting the protein geometries.

To analyze the CD spectra of hemoglobin, we display the transition populations, which are defined as the squares of exciton wave function coefficients. Transition populations of the four CD peaks at 44 000, 48 000, 52 000, and 55 000 cm^{-1} are plotted in Figure 3C. CD peaks at 48 000 and 52 000 cm^{-1} are known to come from exciton splitting of the $\pi\pi^*$ transitions.[34] The 48 000 cm^{-1} transition has polarization parallel to the helical axis, and we see that they are very strong in the helix termini. The 52 000 cm^{-1} transition is polarized perpendicular to the helical axis and is almost evenly distributed across

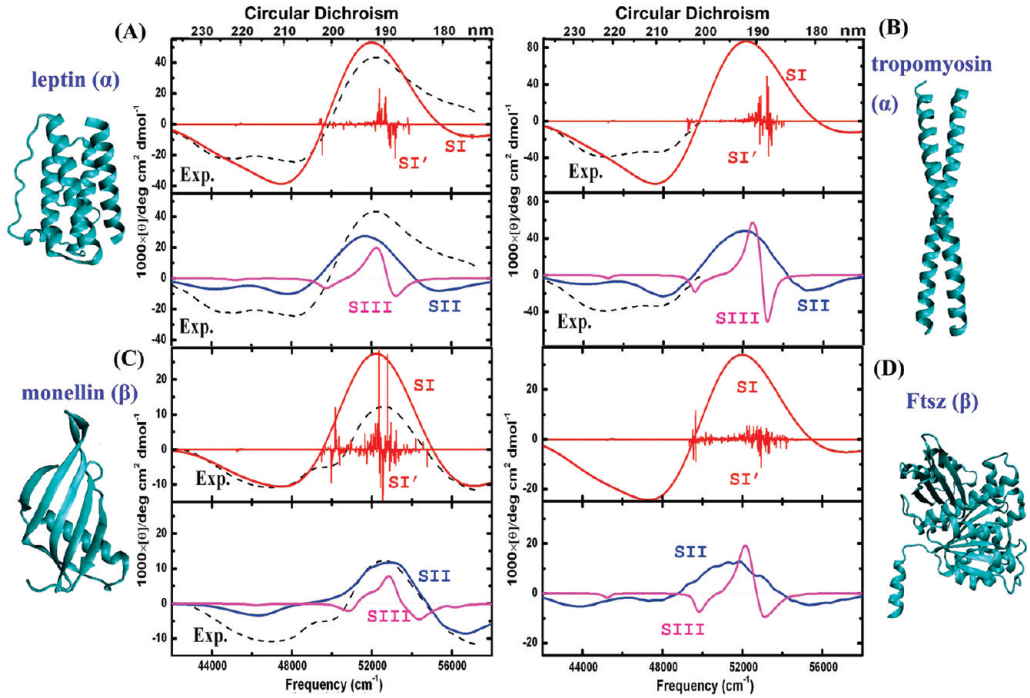

Figure 5. CD spectra of α-helix proteins: leptin (PDB code: 1ax8) and tropomyosin (PDB code: 2d3e), sheet-containing protein: monellin (PDB code: 1mol), and αβ-protein: FtsZ (PDB code: 1fsz), together with X-ray crystal structures. Same labels as in Figure 3.

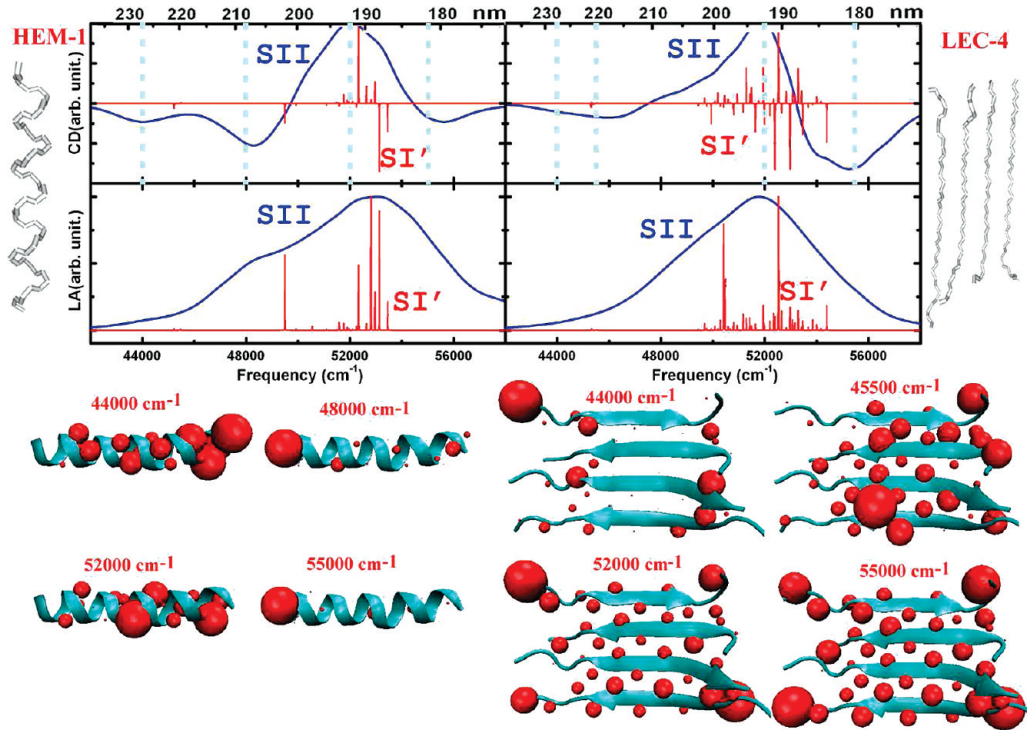

Figure 6. CD and LA spectra of α-helix (HEM-1: fragment of 1hda) and β-sheet (LEC-4: fragment of 1les) model structure. Same labels as in Figure 3.

the protein. The 44 000 cm^{-1} peak comes from the helical regions, whereas the 55 000 cm^{-1} peaks come from the turns. We can now explain the missing CD features in the SI and SIII spectra. This is mainly due to the omission of the electrostatic potential induced by surrounding peptide groups, amino acid side chains, and water molecules. Transition densities of different peptide bonds are affected differently by their surroundings. A single broadening factor for all transitions misses fine structural details of the spectrum. The SII spectra carefully include the electrostatic potentials induced by surrounding molecular groups and capture these finer details.

***β*-Sheet Protein: Lentil Lectin.** Figure 4 shows the simulated CD and LA spectra and the corresponding X-ray crystal structure of lentil lectin (PDB code: 1les), which is a typical β-sheet protein. The experimental CD peaks[31,32] at 45 500, 51 000, and 55 000 cm^{-1} are well reproduced by the simulated SII spectrum of the ensemble of 2000 MD snapshots. Compared with SI spectra based on single conformation, SII spectra provide more

detailed fine structure. The SIII CD spectra are much narrower than SII and experiment, demonstrating the importance of the electrostatic potential due to amino acid side chains. Because the $\pi\pi^*$ exciton splitting is much weaker in sheet structures, we do not observe a CD peak at 48 000 cm^{-1}. The transition populations at frequencies of 44 000, 45 500, 51 000, and 55 000 cm^{-1} are displayed in Figure 4C. The 44 000 cm^{-1} transition occurs at the turn regions. CD peaks at 45 500 and 55 000 cm^{-1} originate from the sheet regions, whereas the 52 000 cm^{-1} peak has contributions from transitions from all peptide groups.

Comparison of Helical and Sheet Proteins. To explore the relationship between the UV spectra and secondary structures, we display in Figure 5 the simulated CD spectra and corresponding X-ray crystal structures of four more proteins: the helical proteins leptin (PDB code: 1ax8) and tropomyosin (PDB code: 2d3e), sheet protein monellin (PDB code: 1mol), and $\alpha\beta$-protein FtsZ (PDB code: 1fsz). In all cases, the simulated SII spectra reproduce the experimental fine structure.[31,33] Compared with SI and SIII, the SII spectra computed from 2000 MD snapshots provide better agreement with experiments.

We now summarize the CD spectra of the three helical protein (hemoglobin, leptin, and tropomyosin) and the two sheet proteins (lentil lectin and monellin). We observe two negative CD peaks at 44 000 and 48 000 cm^{-1} in helical proteins compared with a single strong negative peak at ~45 000−46 000 cm^{-1} in sheet proteins. The negative CD peak at ~55 000−56 000 cm^{-1} is more pronounced in sheet proteins. The $\alpha\beta$-protein FtsZ contains both helix and sheet motifs, and shows the helix feature of two negative peaks at 44 000 and 48 000 cm^{-1} and the sheet feature of intense negative peaks at ~45 000−46 000 and ~55 000−56 000 cm^{-1}.

We have used some model systems to examine the relationships between CD peaks and secondary structures. A helical fragment was extracted from hemoglobin (Pro124-His143 of chain D in 1hda), and a sheet fragment containing four β strands was taken from lectin (4 strands: Thr1-Phe11 of chain C, Val37-Leu46 of chain D, Val60-Val70 of chain C, and Glu158-Ala169 of chain C in 1les). The fragments are denoted as HEM-1 and LEC-4, respectively. The simulated CD and LA spectra are depicted in Figure 6. We see two negative peaks at 44 000 and 48 000 cm^{-1} in the CD spectrum of HEM-1 and two strong negative peaks at ~45 000−46 000 and ~55 000−56 000 cm^{-1} in the CD spectrum of LEC-4. The transition populations corresponding to the CD peaks of HEM-1 and LEC-4 are displayed at the bottom of Figure 6. The 48 000 cm^{-1} peak results from the exciton splitting in helices and is absent in the CD of sheet proteins. Transition populations at 52 000 cm^{-1} are distributed all over both fragments, which is consistent with the observation of intense positive CD signals in both helical and sheet proteins. The transition populations of model systems are consistent with the full proteins, as shown in Figure 3C and Figure 4C. These CD peaks may thus be used to probe the secondary structure.

Conclusions

We have developed a generalized full-space approach combining QM and MM calculations to study the fluctuating effective electronic Hamiltonian in proteins. A large number of structure snapshots were created using MD simulations, some of which were chosen as representative structures of the structural ensemble, and on these we performed a full charge distribution analysis. The EHEF code was used to combine the MM and QM results and to provide a fluctuating trajectory of the excitation parameters. The transition-energy fluctuations of electronic transitions can be evaluated for each selected trajectory point. This allows us to avoid expensive repeated QM calculations and obtain the fluctuating Hamiltonian at the QM level for all of the snapshots. Simulations of UV spectra of proteins in water with fluctuation effects show good agreement with experiment. The bandshapes of CD and LA spectra have been reproduced by simulations without using empirical parameters. The fine structure of the UV spectra has been well described by considering the electrostatic environment.

Acknowledgment. We gratefully acknowledge the support of the National Institutes of Health (grants GM059230 and GM091364) and the National Science Foundation (grant CHE-0745892). J.D.H. thanks the Leverhulme Trust for a Research Fellowship. B.M.B. was the grateful recipient of an Early-Stage Researcher Short Visit award from the Collaborative Computational Project for Biomolecular Simulation. We thank Daniel Healion and Dr. ZhenYu Li for helpful discussions.

References and Notes

(1) Kern, D.; Eisenmesser, E. Z.; Wolf-Watz, M. *Methods Enzymol.* **2005**, *394*, 507–524.
(2) Oskouei, A. A.; Bram, O.; Cannizzo, A.; van Mourik, F.; Tortschanoff, A.; Chergui, M. *Chem. Phys.* **2008**, *350*, 104–110.
(3) Oskouei, A. A.; Bram, O.; Cannizzo, A.; van Mourik, F.; Tortschanoff, A.; Chergui, M. *J. Mol. Liq.* **2008**, *141*, 118–123.
(4) Zhuang, W.; Hayashi, T.; Mukamel, S. *Angew. Chem.* **2009**, *48*, 3750–3781.
(5) Cui, Q.; Karplus, M. *J. Chem. Phys.* **2000**, *112*, 1133–1149.
(6) Hu, H.; Yang, W. T. *THEOCHEM* **2009**, *898*, 17–30.
(7) la Cour Jansen, T.; Dijkstra, A. G.; Watson, T. M.; Hirst, J. D.; Knoester, J. *J. Chem. Phys.* **2006**, *125*, 044312.
(8) Lin, Y.-S.; Shorb, J.; Mukherjee, P.; Zanni, M. T.; Skinner, J. L. *J. Phys. Chem. B* **2009**, *113*, 592–602.
(9) Brahms, S.; Brahms, J. *J. Mol. Biol.* **1980**, *138*, 149–178.
(10) Greenfield, N. J. *Anal. Biochem.* **1996**, *235*, 1–10.
(11) Woody, R. W. *Monatsh. Chem.* **2005**, *136*, 347–366.
(12) Woody, R. W. *J. Chem. Phys.* **1968**, *49*, 4797–4806.
(13) Bulheller, B. M.; Rodger, A.; Hirst, J. D. *Phys. Chem. Chem. Phys.* **2007**, *9*, 2020–2035.
(14) Hirst, J. D. *J. Chem. Phys.* **1998**, *109*, 782–788.
(15) Besley, N. A.; Hirst, J. D. *J. Am. Chem. Soc.* **1999**, *121*, 9636–9644.
(16) Bulheller, B. M.; Hirst, J. D. *Bioinformatics* **2009**, *25*, 539–540.
(17) Hirst, J. D.; Bhattacharjee, S.; Onufriev, A. V. *Faraday Discuss.* **2003**, *122*, 253–267.
(18) Frenkel, Y. *J. Phys. Rev.* **1931**, *37*, 17–44.
(19) Abramavicius, D.; Palmieri, B.; Mukamel, S. *Chem. Phys.* **2009**, *357*, 79–84.
(20) Abramavicius, D.; Palmieri, B.; Voronine, D. V.; Šanda, F.; Mukamel, S. *Chem. Rev.* **2009**, *109*, 2350–2408.
(21) Karlstrom, G.; Lindh, R.; Malmqvist, P.; Roos, B.; Ryde, U.; Veryazov, V.; Widmark, P.; Cossi, M.; Schimmelpfennig, B.; Neogrady, P.; Seijo, L. *Comput. Mater. Sci.* **2003**, *28*, 222–239.
(22) Kurapkat, G.; Kruger, P.; Wolimer, A.; Fleischhauer, J.; Kramer, B.; Zobel, A.; Koslowski, A.; Botterweck, H.; Woody, R. W. *Biopolymers* **1997**, *41*, 267–287.
(23) Luo, Y.; Norman, P.; Ågren, H. *J. Chem. Phys.* **1998**, *109*, 3589–3595.
(24) Frisch, M. J.; et al. *Gaussian 03*, revision C.02; Gaussian, Inc.: Wallingford, CT, 2004.
(25) Krueger, B. P.; Scholes, G. D.; Fleming, G. R. *J. Phys. Chem. B* **1998**, *102*, 9603–9604.
(26) Madjet, M. E.; Abdurahaman, A.; Renger, T. *J. Phys. Chem. B* **2006**, *110*, 17268–17281.
(27) MacKerell, A. D., Jr.; et al. *J. Phys. Chem. B* **1998**, *102*, 3586–3616.
(28) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983**, *79*, 926–935.
(29) Phillips, J.; Braun, R.; Wang, W.; Gumbart, J.; Tajkhorshid, E.; Villa, E.; Chipot, C.; Skeel, R.; Kalé, L.; Schulten, K. *J. Comput. Chem.* **2005**, *26*, 1781–1802.
(30) Besley, N. A.; Oakley, M. T.; Cowan, A. J.; Hirst, J. D. *J. Am. Chem. Soc.* **2004**, *126*, 13502–13511.
(31) Bulheller, B. M.; Miles, A. J.; Wallace, B. A.; Hirst, J. D. *J. Phys. Chem. B* **2008**, *112*, 1866–1874.
(32) Lees, J. G.; Miles, A. J.; Wien, F.; Wallace, B. A. *Bioinformatics* **2006**, *22*, 1955–1962.
(33) Bulheller, B. M.; Rodger, A.; Hicks, M. R.; Dafforn, T. R.; Serpell, L. C.; Marshall, K.; Bromley, E. H. C.; King, P. J. S.; Channon, K. J.; Woolfson, D. N.; Hirst, J. D. *J. Am. Chem. Soc.* **2009**, *131*, 13305–13314.
(34) Moffitt, W. *J. Chem. Phys.* **1956**, *25*, 467–478.

JP101980A