DEVELOPMENT OF MOLECULAR DYNAMICS FORCE FIELD OF YOPRO-1 AND DEEP LEARNING MODELS FOR PROTEIN CLASSIFICATION

By

Chi Jin

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for degree of

Chemistry—Doctor of Philosophy

2019

ABSTRACT

DEVELOPMENT OF MOLECULAR DYNAMICS FORCE FIELD OF YOPRO-1 AND DEEP LEARNING MODELS FOR PROTEIN CLASSIFICATION

By

Chi Jin

Cyanine dyes, such as Oxazole yellow (YOPRO), are almost non-fluorescent in water but their fluorescence is greatly enhanced after intercalation in double-stranded DNA, providing the basis of DNA concentration assays. The rationale for this property is the flexibility difference of the conformations of the molecule in different environments, mainly attributed to the linker dihedral rotations.

We compared two methods for deriving the specific dihedral force field on the linker of YOPRO, namely by modifying the AMBER generated force field (GAFF) and by using the IPolQ fitting protocol. There are two dihedral angles and the IPolQ method showed that their potential surfaces are coupled. Thus, going beyond the GAFF approach, coupled dihedral surfaces were obtained for the ground S0 and first excited S1 electronic states. Molecular Dynamics (MD) simulations of YOPRO were carried out in water and intercalations using these force field models. The MD simulations started at the minima of the S0 state vertically excited to the S1 state. The contrast between YOPRO conformational relaxation on the S1 surface in water and when intercalated provided the non-radiative relaxation pathways relevant to fluorescence decay and explain the differences in quantum yield.

For the second topic, we investigated a number of deep machine learning (ML) models for protein family classification. We used one dimensional sequence and three dimensional secondary structural information of proteins as the input for training the neural network models. The results show that deeper convolutional networks of the Long Short Term Memory (LSTM) variety significantly enhanced the prediction accuracies compared to less sophisticated models. The addition of the secondary structural information greatly increases the testing accuracies with the training data size remaining the same. Proteins belonging to different families can be successfully distinguished using these methods.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1 Introduction

1.1 General Introduction

Oxazole yellow (YOPRO), shown in Figure 1.1.1, is a cyanine-based dye widely used in commercial DNA concentration assays.[1, 2] During gel electrophoresis for DNA detection, it forms a stable complex with the DNA to avoid its removal.[3, 4] Furthermore, it is also used to monitor the formation and break up of protein-DNA complexes,[5, 6] due to its high binding affinity to double-stranded DNA. At low concentrations, the preferred binding mode of YOPRO is intercalation between paired bases[7], whereas at higher dye-to-DNA ratios, other binding modes on the surface, like external electrostatic binding or minor groove binding can occur.

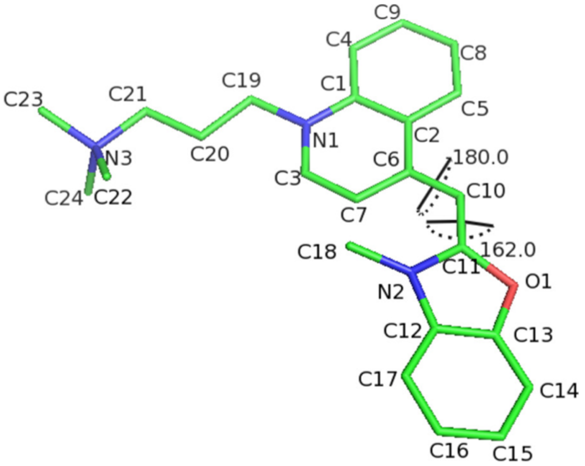

Figure 1.1.1 Structure of YOPRO (without hydrogen atoms)

YOPRO is almost non-fluorescent in water but its fluorescence is enhanced ~1800 fold after intercalation in double-stranded DNA, providing the basis of DNA concentration assays[1, 8-10]. The rationale for this difference is that in water rotation around the linker connecting the benzoxazole and quinoline rings (Figure 1.1.1) lets the excited molecule decay through nonradiative relaxation pathways, thus quenching the fluorescence. However, intercalation into double-stranded DNA enhances the constraints on the rotational motion around the linker, therefore eliminating this pathway and results in the observed intense fluorescence.

In this work, we first carried out molecular dynamics (MD) simulations to explore the conformational sampling of YOPRO in the different environments as discussed above and constructed free energy surfaces (potentials of mean force) along the dihedral angles using modified Amber generated force field models. The simulation results and defects of these force field models are discussed in Chapter 2. Then in Chapter 3, we introduce a more accurate dihedral force field by using the implicitly polarized charge method (IPolQ)[11] for both the ground (S0) and the first excited electronic states (S1). Then, MD simulations using the new force fields are carried out on both electronic surfaces for water solvated and intercalated dye. Finally, a steepest decent algorithm was implemented to find the non-radiative relaxation pathway on the S1 free energy surfaces for both YOPRO in water and intercalated. The distinct energy barrier found on the intercalated path that is absent in the water solvated path supports the hypothesis that constraints on linker between the two ring systems are responsible for the large difference in fluorescence intensity in different environments.

Chapter 4 is devoted to the different topic of protein family classification[12] using Machine Learning[13]. Proteins are classified into families based on evolutionary ancestry. A given protein family has similar three dimensional structure and function. In our investigations we studied if a neural network can be trained to discriminate between protein families based on sequence alone, and based on sequence and corresponding structure. That sequence alone can discriminate between different families indicates that sequence does encode protein structure and function.

We investigated different neural network models for predicting protein families and found that the long short-term memory networks (LSTM)[14] show the highest performance among all the models studied. Different methods for extracting protein structural information as the input data, by mapping the extracted structural information into protein family labels, were considered.

Because there are many more sequences known than their structures and because machine learning is data intensive there is a tradeoff between the sequence and sequence plus structure approaches.

All in all, the LSTM neural network approach is shown to be capable of good discrimination in binary comparison of pairs of protein families.

1.2 Molecular Dynamics Simulation

1.2.1 General Description

Molecular dynamics[15, 16] (MD) is a method that simulates the motion of molecules in a closed system using by computationally integrating Newton's equations of motion. The inter and intramolecular motions are driven by a force field that is predesigned to approximate physical laws. Therefore, MD simulations can provide dynamic information of a molecular system at atomic level. For instance, by examining the resulting trajectory of a simulation which is a sample in an ensemble, one can evaluate the average system properties such as energy, entropy and free energy by calculating their time averages.

The algorithms that implement MD simulations are to answer one fundamental question: if the current configuration of the system is given, what is for next moment? By answering it repeatedly, MD propagates the system in time following Newton's equation of motion:

$$\ddot{r} = \frac{F}{m} \tag{1.2.1}$$

Specifically, three steps are involved for MD to propagate the system: first, one needs to initialize the positions and velocities to all the atoms in the system. For most biological systems, MD starts out with the crystal structure of the molecules of interest. To conduct simulation in solvent, either implicit solvent is added as a continuous media or explicit solvent are added by putting the system into an equilibrated box of solvent molecules. Then the velocities of the system are initialized with a random seed generated from a Maxwell distribution[17]. Secondly, the forces are evaluated according to the force field. Finally, an integrator is used to predict the configuration for the next moment given a certain time step. The last two steps are repeated until the preset number of step are completed.

1.2.2 Force Field

A MD force field is a set of functional forms and parameters that are used to simulate the potential energy change due to the intramolecular interactions and intermolecular motions. The force field parameters is developed to fit the experimental data or quantum calculation results into the following terms: bond stretches and vibrations, dihedral torsions, Lennard-Jones and electrostatic potentials. A number of force fields have been developed for biological systems such as the AMBER[18], CHARMM[19] and GROMOS[20] force fields. As an example, equation 1.1.2 shows the functional form of the AMBER internal bonded force field:

$$V(r^N) = \sum_{bonds} k_b (l - l_0)^2 + \sum_{angles} k_a (\boldsymbol{\theta} - \boldsymbol{\theta}_0)^2 + \sum_{torsions} \sum_{n} \frac{1}{2}[1 + \cos(n\boldsymbol{\omega} - \boldsymbol{\gamma})] \qquad (1.2.2)$$

The first term on the right is the bond stretching term as the function of the bond length l with the equilibrium bond length at l_0 and force constant equal k_b . The second term is the bond bending term as the function of the bond angle $\boldsymbol{\theta}$ with the equilibrium bond angle at $\boldsymbol{\theta}_0$ and force constant equal k_a . The last one is the dihedral torsional term as a function of $\boldsymbol{\omega}$ with the multiplicity n and $\boldsymbol{\gamma}$ as the phase shift .

1.2.3 Integrator

Many integrators have been designed, one of which is the leap-frog algorithm[21]:

$$\begin{aligned} r(t+\Delta t) &= r(t) + v(t-\frac{\Delta t}{2})\Delta t \\ v(t+\frac{\Delta t}{2}) &= v(t-\frac{\Delta t}{2}) + a(t)\Delta t \end{aligned} \qquad (1.2.3)$$

where t is the time, and r(t), v(t) and a(t) are the coordinates, the velocity and the acceleration of the particles. The integration step, Δt, should be chosen small enough to catch the fastest motion which is the hydrogen related bond vibrations. To speed up the MD, the SHAKE algorithm[17] is usually used to constrain these bonds to increase the step size to the scale of 1fs. This should be a good approximation since these bond vibration motions will be averaged out in a normal simulation scale.

1.2.4 Period Boundary Condition

Period Boundary Conditions (PBC) are used to simulate an infinite system with a finite size box by representing it as a repeated array of the box (Figure 1.2.1). Certainly an artificial periodicity is imposed on the system simulated. In conjunction with using PBC, the long range interactions can be calculated using the Ewald Lattice Sum[17] implemented by the Particle Mesh Ewald method[22].

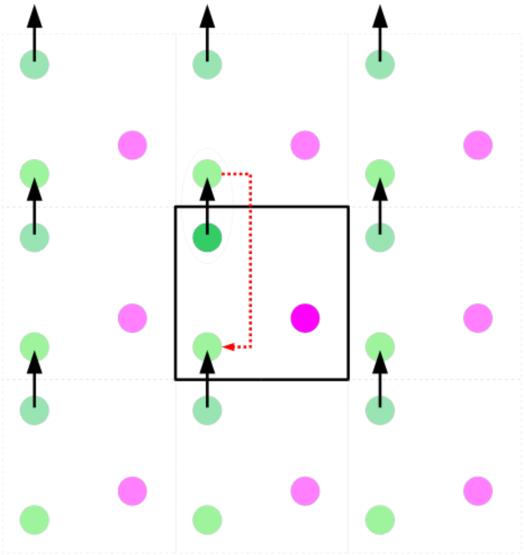

Figure 1.2.1 Illustration of the Periodic Boundary Condition. The center black box is the primary box and all the others are the copies of it. Once the dark green particle in the primary box goes out, the light green one in the bottom copy enters the box to compensate. From https://en.wikipedia.org/wiki/Periodic_boundary_conditions

1.2.5 Temperature and Pressure Control

The Newton's equation (equation 1.2.1) conserves the total energy of the system simulated. Therefore, without temperature control, the system is simulated in a microcanonical ensemble with fixed number of particles (N), volume(V) and energy(E). However, the normal experimental conditions usually allows the total energy to fluctuate but fix temperature and the pressure of the system. Therefore, we need to control the temperature and pressure.

Several methods have been developed for this problem[23], one of them is the Berendsen[24] method. It couples the simulated system with an external bath of the temperature T_{ref} . For any particle i, Berendsen et al proposed the modification on the velocity:

$$\ddot{r}_i = \frac{F_i}{m_i} + \frac{1}{2\tau}(\frac{T_{ref}}{T} - 1)\dot{r}_i \tag{1.2.4}$$

where τ is the coupling time and usually τ=0.2ps is used. T is the instantaneous temperature, which is equal to $\frac{2}{3k_B N}\sum_{i=1}^{N}\frac{1}{2}m_i\dot{r}_i^2$, where N is the total number of particles k_B is the Boltzmann constant. The extra term added in equation 1.2.3 serves as a frictional force as a feedback mechanism to scale T back to T_{ref} . The velocity is therefore scaled by a factor $\sigma = \sqrt{1 - \frac{\Delta t}{\tau}\frac{T - T_{ref}}{T}}$, where Δt is the MD step size.

The instantaneous pressure P is computed by

$$P = \frac{2}{3V}(K - Virial) \tag{1.2.5}$$

where K is the kinetic energy of the system and Virial is the virial function[17]. The pressure of a n isotropic system is constrained by a factor of γ:

$$\gamma = (1 - \frac{\Delta t}{\beta}\frac{P - P_{ref}}{P})^{-\frac{1}{3}} \tag{1.2.6}$$

where P_{ref} is the reference pressure. For an anisotropic box, the pressure control is complicated where a pressure tensor has to be used[24]. Given the simplicity of the implementation of the Berendsen method, however, it doesn't lead to any known ensemble. To guarantee an NPT ensemble, methods such as the Langevin dynamics need to be used[25].

1.2.6 Long-range Interactions

The potential energy of the non-covalent interactions is usually expressed as a sum of the pair-wise electrostatic and Lennard-Jones contributions:

$$V_{nb} = \sum_{i,j} (\frac{q_i q_j}{r_{ij}} + \frac{A_{ij}}{r_{ij}^{12}} + \frac{B_{ij}}{r_{ij}^{6}}) \qquad (1.2.7)$$

The first term on the right, which represents the electrostatic contribution, has a form of $\frac{Kq_i q_j}{r^p}$ with p =1. It is known that a summation of long-range interaction terms of this form with $p \leq 3$ is a conditionally convergent sum[26]. The easiest way to solve the convergence problem is to use a cut-off method, which simply ignores the interactions beyond some chosen cutoff distance. But later more accurate methods were developed, such as the Ewald summation[17]. It divides the calculation of summation into two parts that are in the real space and the reciprocal space, respectively, which converge rapidly given the system is charge neutral. The original reciprocal summation was implemented in $O(N^2)$, which is still too expensive for large systems. Particle Mesh Ewald (PME) method[22] makes interpolation of the reciprocal structure factor in the lattice points and uses Fast Discrete Fourier Transform (FDFT) to reduce the time cost to O(NlogN).

1.3 Weighted Histogram Analysis Method

1.3.1 Umbrella Sampling

The current affordable time scale for MD simulations for a normal-sized protein in explicit solvents is limited to only tens of nanoseconds. Therefore, when there are large barriers in the complex potential energy surface of the molecule of interest, MD may not be able to sample the configuration space properly. To compensate for this defect, umbrella sampling method is often used[23]. In an umbrella sampling, a restraint potential, which is usually a parabolic function of the coordinates, is added to the original potential surface to force the system to stay around the restraint coordinate and make a biased energy surface. This modified system is called a window. Multiple restraint potentials centered at different locations can be applied to make multiple windows, which are sampled by MD one after the other. Finally, the resulting window biased probability data are combined (figure 1.3.1) and unbiased using the Weighted Histogram Analysis method(WHAM)[27, 28] to construct a potential of mean force (PMF)[29, 30] of the whole sampling space.

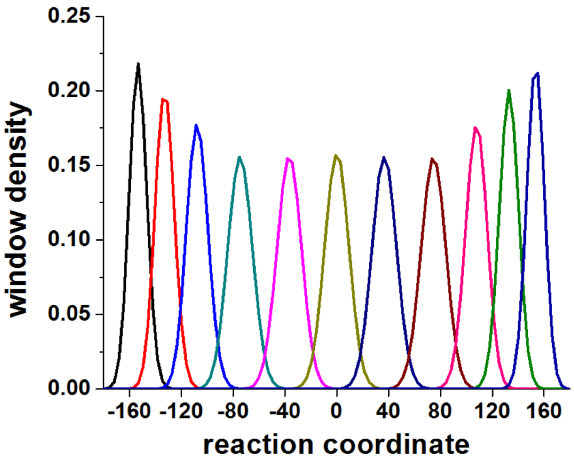

Figure 1.3.1 Overlapping of the probability peaks of all the windows sampling the angle space of a bond angle[31].

1.3.2 Unbiasing Procedure

Suppose the original potential energy surface $U_0(\eta)$ and the restraint potential of the ith window W_i are functions of the reaction coordinate $\eta = \eta(\mathbf{R})$, where $\mathbf{R}$ is the set of atom coordinates. Thus, the biased and unbiased probability distributions are:

$$\rho_i^{(b)}(\eta) = \frac{e^{-(W_i(\eta)+U_0(\eta))\beta}}{\sum_{\eta} e^{-(W_i(\eta)+U_0(\eta))\beta}} \tag{1.3.1}$$

and

$$\rho_i^{(u)}(\eta) = \frac{e^{-U_0(\eta)\beta}}{\sum_{\eta} e^{-U_0(\eta)\beta}} \tag{1.3.2}$$

respectively. By plugging equation 1.3.2 into equation 1.3.1, we can obtain:

$$\rho_i^{(u)}(\eta) = \rho_i^{(b)}(\eta) e^{\beta W_i(\eta)} \frac{\sum_{\eta} e^{-(W_i(\eta)+U_0(\eta))\beta}}{\sum_{\eta} e^{-U_0(\eta)\beta}} \tag{1.3.3}$$

We now define the free energy of a certain of the ith window f_i using the third term on the right of equation 1.3.3:

$$e^{-\beta f_i} \equiv \frac{\sum_{\eta} e^{-(W_i(\eta)+U_0(\eta))\beta}}{\sum_{\eta} e^{-U_0(\eta)\beta}} = \sum_{\eta} \rho_i^{(u)}(\eta) \times e^{-W_i(\eta)\beta} \tag{1.3.4}$$

Now we combine the unbiased probability $\rho^{(u)}(\eta)$ of all the windows to approximate the global unbiased probability $\rho_0(\eta)$, and use $\rho_0(\eta)$ to replace $\rho^{(u)}(\eta)$ in equation 1.3.4:

$$e^{-\beta f_i} = \sum_{\eta} \rho_0(\eta) \times e^{-W_i(\eta)\beta} \tag{1.3.5}$$

where

$$\rho_0(\eta) = \sum_{i=1}^{N} p_i(\eta) \times \rho_i^{(u)}(\eta) \tag{1.3.6}$$

where $p_i(\eta)$ is the weight of the ith windows, and it is normalized:

$$\sum_{i=1}^{N} p_i(\eta) = 1 \tag{1.3.7}$$

In order to minimize the statistical error on the global probability distribution[32], we differentiate the standard deviation with respects to the weights and set them to 0:

$$\frac{\partial(\sigma^2[\rho_0(\eta)])}{\partial p_i} = 0 \tag{1.3.8}$$

to derive:

$$\rho_0(\eta) = \frac{\sum_{i=1}^{N} n_i \rho_i^{(b)}(\eta)}{\sum_{j=1}^{N} n_j e^{-\beta(W_j(\eta)-f_j)}} \tag{1.3.9}$$

Note that n_i is the number of data points in ith window.

The resulting $\rho_0(\eta)$ then goes back to equation 1.3.5 to start the next iteration. So on and so forth until the difference of the f value for all the windows obtained from two iterations converges to some given tolerance. The final $\rho_0(\eta)$ is then used to construct the PMF surface.

1.4 Implicitly Polarized Charge Method

The simulation of the solute molecule of interest relies on the construction of accurate force field parameters[33]. The Implicitly Polarized Charge method (IPolQ) derives the partial charge set of the molecule in solution as the first step to obtain these potentials from ab initio quantum calculations[11]. To start with, the solute molecule is solvated in a water box, and it is kept fixed. A 10 ns conformational sampling of the water solvent at 450K is carried out to collect a set of equilibrated water configurations. The solvent charge density and the electrostatic field of these water configuration is calculated as the reaction field in Figure 1.4.1.

For the second step, the MD simulation of the solute molecule is performed from each equilibrated water configuration with the water solvent fixed to obtain a set of time average solute configurations. The fixed locations of the water solvent later will serve to create a field of point charges surrounding the solute.

Once the MD simulation of the solute is finished, the quantum chemistry program orca[34] is used to calculate the single point energy and electrostatic potential of each solute configuration in vacuum for charge fitting. The solvent charge density calculated in the first step is taken as a perturbation for the electrostatic potentials of the solute in water. The final IPolQ charge set was derived from the perturbed electrostatic potential set using the Restrained Electrostatic Potential (RESP) model[35].

The IPolQ charges set was used to replace the original guess to start the next iteration of charge fitting. Once the fitted charge set is converged, the charge fitting process is finished. The IPolQ charge set then goes to the force constant fitting section.

The architecture of whole procedure is show as below:

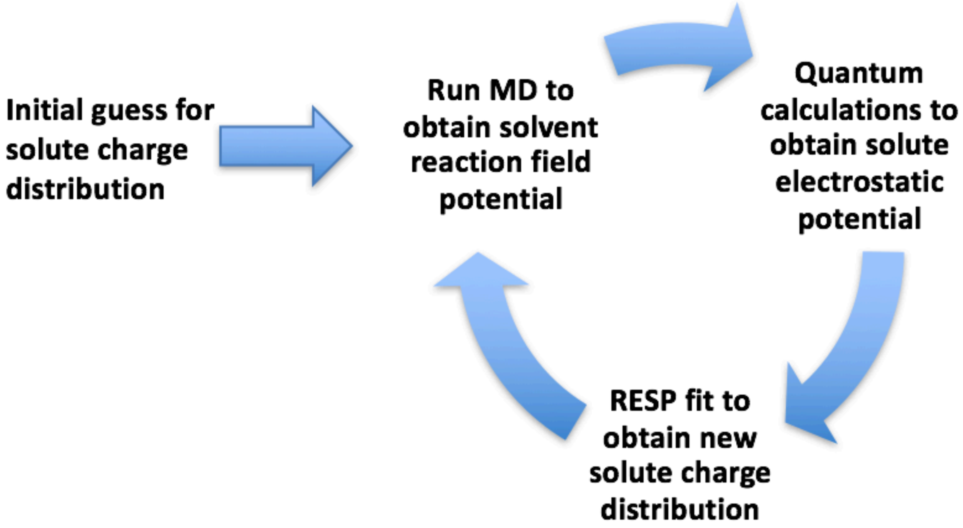
Initial guess for solute charge distribution
Run MD to obtain solvent reaction field potential
Quantum calculations to obtain solute electrostatic potential
RESP fit to obtain new solute charge distribution

Figure 1.4.1 Work cycle of the IPolQ Method

1.5 Force Constant Fitting

The partial charge set of the molecule of interest obtained by the IPolQ method in Section 1.4 is used to sample a set of conformations in vacuum by systematically varying the investigated degrees of freedom (one or two) for which the force constants are to be fitted and optimizing the rest part of the molecule.

For each conformation, the single point energy Q and the classical potential energy V from the AMBER Generalized Force Field[18] with the fitted terms excluded were calculated. Therefore, the pure contribution of the fitted terms is computed by subtracting V from Q to construct a "subtracted" potential surface along them.

If we have two terms to fit and the resulting potential surface shows a strong coupling of them then, for each conformer, we fit the force constant parameters of one of them (D1) with the other(D2) fixed at a set of preset values. For any fixed D2 value, a Fourier expansion of the potential energy of D1 with four terms (equation 1.2.2) is used. For instance, if both D1 and D2 are dihedrals, we use the following equation to express the torsional energy $V_{(D1|D2)}(x)$ of D1 (x) at a fixed D2 value:

$$\begin{aligned}
&V_{(D1|D2)}(x) = a_0 + \frac{a_2}{2}(1+\cos(2x-b_2)) \\
&+\frac{a_4}{4}(1+\cos(4x-b_4)) \\
&+\frac{a_6}{6}(1+\cos(6x-b_6)) \\
&+\frac{a_8}{8}(1+\cos(8x-b_8))
\end{aligned} \tag{1.5.1}$$

where the a_n and b_n values are the fitted force field parameters for the energy barrier and phase shift, respectively.

The whole procedure is repeated with the roles of D1 and D2 interchanged; thus, each conformation has a fitted force field parameter set for its corresponding window in the later umbrella samplings.

1.6 Fluorescence Theory

An atom absorbs a photon when resonance happens, namely the oscillations of the light wave are coupled with the oscillation of the electrons of the atom, causing the electron distribution to "reshape" and conversion of the low electronic state to higher electronic state, which the has more nodes on the orbital surface[36]. The energy gap between the two state is equal to the energy of the photon: $\Delta E = h\nu$.

In a molecule, this electronic transition is coupled with nucleic motions such as vibrations, rotations. According to the Franck-Condon principle, since the electronic motions are much faster than the nuclear motions, the molecule maintains its nuclear structure after the transition, therefore the transition often ends at higher vibrational states rather than the ground state as indicated by the end of the blue arrow in Figure 1.6.1. The excited molecule usually quickly relaxes to the lowest vibrational state of the excited electronic state through non-radiative relaxations by dissipating heat to the surroundings and from there decays to the ground electronic state as shown with the green arrow in Figure 1.6.1, where a photon, known as the fluorescence is emitted.

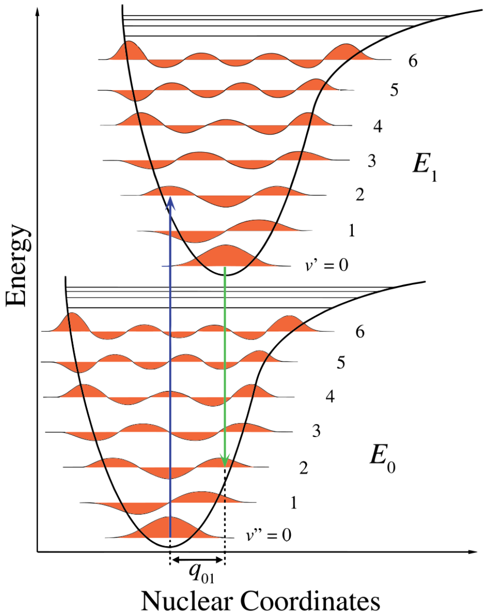

Figure 1.6.1 Fluorescence Theory. https://en.wikipedia.org/wiki/Franck%E2%80%93Condon_principle

In addition to fluorescence, the excited molecule may also decay to the ground electronic state via non-radiative transitions at critical nuclear configurations, r_c (Figure 1.6.2), which correspond to the minima on the excited surface. In the vicinity of r_c, ψ(ground) and ψ(excited) are mixed, especially when the energy gap ΔE is large and the surface jump is "strongly avoided", which leads the electron to stay on the same adiabatic surface. [36] On the other hand, when the energy gap is small, the surface jump becomes possible near r_c so the electron can therefore go down to the ground state surface .

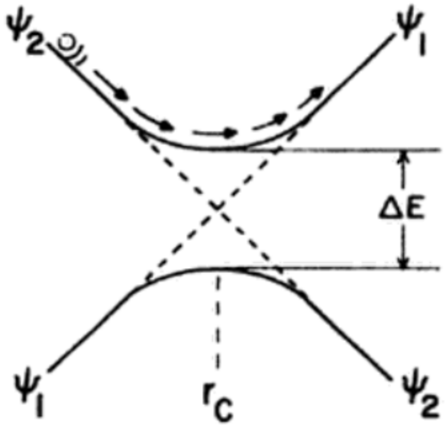

Figure 1.6.2 Avoided crossing. (Turro P157, Fig 6.3c)

Another type of non-radiative transitions happen at conical intersections (CI), where the S1 and S0 surfaces intersects(Figure 1.6.3). In this case, there's no such region where ψ(ground) and ψ(excited) are mixed except the CI point.

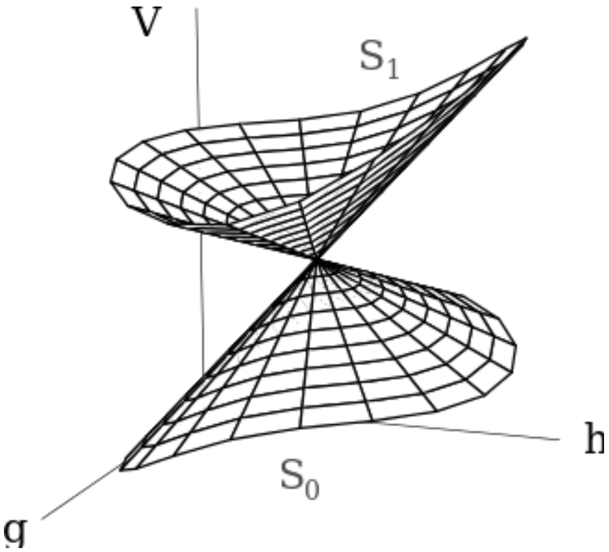

Figure 1.6.3 Conical Intersection. https://commons.wikimedia.org/wiki/File:Conical_intersection.svg

The basis of DNA concentration assays using cyanine dyes like YOPRO relies on the fact that these dyes are almost non-fluorescent in aqueous solutions, but intensely fluorescent when intercalated in DNA.[1, 9, 10, 37-40] This results from the different configurational distributions in the two environments. In solution, twisting around the linker between the two ring systems leads to geometries that let the excited dye access nonradiative decay through avoided crossing or CI, quenching the fluorescence. However, when intercalated, the environment rigidity increases the constraints on the rotational motion around the linker by forming energy barriers on the pathway, resulting in fluorescence enhancements of greater than 1,000 fold.[7, 41] This mechanism has been suggested broadly in the context of molecular photochemistry arising from increasing medium viscosity[42]; for example, the molecular motions of the stilbenes in solution versus rigid environments. [36, 43] It has been studied that the quantum yield of fluorescence of the stilbenes depends on rotational isomerization which is inhibited by the high viscosities.[44] Moreover, in the first excited singlet state of stilbenes, there has been found some vibrational modes that are responsible for the low quantum yield at low viscosities. The population of these modes are all decreased at high viscosities where an increase in the quantum yield is observed. These motion are absent in rigid nonplanar molecules even at low viscosities.[43, 44] For YOPRO, this scenario has also been inferred from the similarity between the activation energies for the temperature dependence of non-radiative decay processes and for viscous flow[38].

Similar results has also been observed for the dyes such as DPTox, T3ox[42] and YO[38] in a viscous solution, where the rotational mobility around the linker is decreased and the corresponding activation energy for the temperature-dependent nonradiative decay is similar to that for the intercalation. Based on these, a series of hemicyanine derivatives are used as sensors for the viscosity in solutions or biological fluids in vivo. By measuring the lifetime, intensity and anisotropy of the fluorescence, the viscosity of the cell membranes and the cytoplasm can be determined for biological researches of intracellular mass and signal transport, reactive metabolites and biomacromolecules interactions and diffusions, and clinical diagnoses. By making choice of the heterocycles on the hemicyanine dyes, different levels of sensitivity can be achieved.[39] In contrast, understanding this mechanism helps to increase the quantum yield of many chromophores used as noninvasive markers. By binding them to rigid macromolecular matrixes to restrict the twisting motions, their fluorescence intensities are raised by several orders of magnitude.[41] A good example is the green fluorescent protein (GFP), which is used to mark gene expression and protein localization. Restraining the twisting degrees of freedom of its chromophores by tightly fixing the inside the protein exhibits a quantum yield of ~0.8,[40] and the fluorescence emission can be reduced by protease digestion or heat denaturation, when the chromophores regain their flexibility. This is analogue to the intercalation of cyanine dye molecules like YOPRO in the double-stranded DNA.

1.7 Kramers Theory

In 1940, Kramers developed a model describing the reaction rate of thermally activated barrier crossing processes.[45, 46] As shown in Figure 1.7.1, a classical particle of mass M is moving in a asymmetric double-well potential U(x), where the two local minima at x_a and x_c on the reaction coordinate axis X designate the initial and final states, and the local maximum at x_b gives the energy barrier of E_b^+. We now derive the escaping rate of the particle over the potential barrier in consequence of Brownian motion.

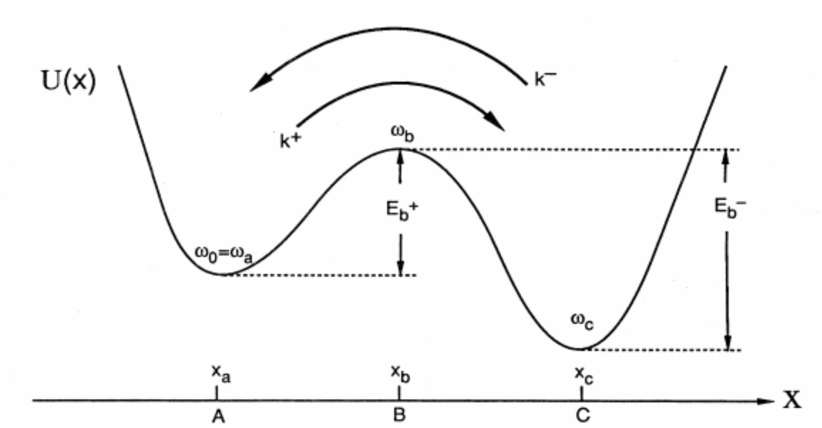

Figure 1.7.1 Potential U(x) with two metastable states A and C. Escape occurs via the forward rate k^+ and the backward rate k^- respectively, and E_b^+ is the corresponding activation energy (Reaction Rate Theory -- 50 Years after Kramers, FIG 3).

Given the Langevin Equation,

$$M\ddot{x} = -\frac{\partial U(x)}{\partial (x)} - \gamma \dot{x} + \eta \tag{1.7.1}$$

where γ is the fiction coefficient, and η is fluctuating force.

In the overdumped region where the inertial term (left hand side of equation 1.5.1) is neglected, by reordering the equation, we have:

$$\gamma\dot{x} = -\frac{\partial U(x)}{\partial(x)} + \eta \tag{1.7.2}$$

The Fokker-Planck equation for the probability density P(x,t) corresponding to this Langevin equation, is[47]

$$\begin{aligned}\frac{\partial P(x,t)}{\partial t} &= \frac{\partial}{\partial x}[\frac{1}{\gamma}\frac{\partial U}{\partial x}P(x,t)] + D\frac{\partial^2 P(x,t)}{\partial x^2}\\ &= \frac{\partial}{\partial x}[\frac{1}{\gamma}\frac{\partial U}{\partial x}P(x,t) + D\frac{\partial P(x,t)}{\partial x}]\\ &= -\frac{\partial J}{\partial x}\end{aligned} \tag{1.7.3}$$

where J is known as the current density:

$$\begin{aligned}J &= -\frac{1}{\gamma}\frac{\partial U}{\partial x}P(x,t) - D\frac{\partial P(x,t)}{\partial x}\\ D\gamma &= k_B T\end{aligned} \tag{1.7.4}$$

with *D* a diffusion coefficient.

Therefore, J can be rewritten as:

$$\begin{aligned}J &= -\frac{D}{k_B T}\frac{\partial U}{\partial x}P(x,t) - D\frac{\partial P(x,t)}{\partial x}\\ &= -De^{-\frac{U(x)}{k_B T}}\frac{\partial(e^{\frac{U(x)}{k_B T}}P)}{\partial x}\end{aligned} \tag{1.7.5}$$

Hence

$$\frac{\partial(e^{\frac{U(x)}{k_B T}}P)}{\partial x} = -\frac{J}{D}e^{\frac{U(x)}{k_B T}} \tag{1.7.6}$$

By integrating both sides from A to C, we get:

$$e^{\frac{U(x)}{k_B T}}P\Big|_A^C = -\frac{J}{D}\int_A^C e^{\frac{U(x^{'})}{k_B T}}dx^{'} \tag{1.7.7}$$

By assuming P(x=C) is very small, because the activation energy is large compared with k_BT, we get:

$$
\begin{aligned}
&-e^{\frac{U(a)}{k_BT}}P(x=A)=-\frac{J}{D}\int_A^C e^{\frac{U(x')}{k_BT}}dx' \\
&\Rightarrow J=\frac{De^{\frac{U(a)}{k_BT}}P(x=A)}{\int_A^C e^{\frac{U(x')}{k_BT}}dx'}
\end{aligned}
\tag{1.7.8}
$$

We now want to evaluate the escape rate r. First, if the barrier is high then we have approximate equilibrium, which makes J in equation 1.7.5 close to 0, therefore:

$$
\begin{aligned}
&\frac{\partial(e^{\frac{U(x)}{k_BT}}P)}{\partial x}\rightarrow 0 \\
&\Rightarrow P(x)=\frac{P(A)e^{\frac{U(A)}{k_BT}}}{e^{\frac{U(x)}{k_BT}}}=P(A)e^{-\frac{U(x)-U(A)}{k_BT}}
\end{aligned}
\tag{1.7.9}
$$

The probability of finding the particle in the well A is therefore:

$$
p=\int_{A-\Delta}^{A+\Delta}P(x)dx=P(A)e^{\frac{U(A)}{k_BT}}\int_{A-\Delta}^{A+\Delta}e^{\frac{-U(x)}{k_BT}}dx
\tag{1.7.10}
$$

where Δ is the size of the well A. The integrand is peaked at x=A, hence:

$$
\begin{aligned}
&\int_{A-\Delta}^{A+\Delta}e^{\frac{-U(x)}{k_BT}}dx=\int_{A-\Delta}^{A+\Delta}e^{\frac{-[U(A)+U'(A)(x-A)+\frac{1}{2}U''(A)(x-A)^2+...]}{k_BT}}dx \\
&\approx e^{\frac{-U(A)}{k_BT}}\times\int_{A-\Delta}^{A+\Delta}e^{\frac{0+\frac{1}{2}U''(A)(x-A)^2}{k_BT}}dx \\
&=e^{\frac{-U(A)}{k_BT}}\times\int_{A-\Delta}^{A+\Delta}e^{\frac{U''(A)(x-A)^2}{2k_BT}}dx \\
&\approx e^{\frac{-U(A)}{k_BT}}\times(\frac{2\pi k_BT}{U''(a)})^{\frac{1}{2}}
\end{aligned}
\tag{1.7.11}
$$

So

$$p = P(A)e^{\frac{U(A)}{k_BT}} \times e^{-\frac{U(A)}{k_BT}} \times (\frac{2\pi k_B T}{U^{''}(A)})^{\frac{1}{2}} = P(A)\times(\frac{2\pi k_B T}{U^{''}(A)})^{\frac{1}{2}} \qquad (1.7.12)$$

for the denominator of the current J in equation 1.7.8, similarly, we have

$$\begin{aligned}
&\int_A^C e^{\frac{U(x^{'})}{k_BT}} dx^{'} \approx \int_{B-\Delta}^{B+\Delta} e^{\frac{[U(b)+U^{'}(b)\times(b-x)+\frac{1}{2}U^{''}(b)\times(b-x)^2]}{k_BT}} dx^{'} \\
&= e^{\frac{U(b)}{k_BT}} \int_{B-\Delta}^{B+\Delta} e^{\frac{\frac{1}{2}U^{''}(b)\times(b-x)^2}{k_BT}} dx^{'} \qquad (1.7.13) \\
&= e^{\frac{U(b)}{k_BT}} (\frac{2\pi k_B T}{U^{''}(B)})^{\frac{1}{2}}
\end{aligned}$$

so

$$J = \frac{De^{\frac{U(A)}{k_BT}} P(A)}{e^{\frac{U(B)}{k_BT}} (\frac{2\pi k_B T}{U^{''}(B)})^{\frac{1}{2}}} \qquad (1.7.14)$$

Finally, the Kramers flux over probability escape rate is

$$\begin{aligned}
r &= \frac{J}{p} = \frac{\dfrac{De^{\frac{U(A)}{k_BT}} P(A)}{e^{\frac{U(B)}{k_BT}} (\frac{2\pi k_B T}{U^{''}(B)})^{\frac{1}{2}}}}{P(A)\times(\frac{2\pi k_B T}{U^{''}(A)})^{\frac{1}{2}}} \\
&= \frac{D}{2\pi k_B T}[U^{''}(A)U^{''}(B)]^{\frac{1}{2}} e^{\frac{U(A)-U(B)}{k_BT}} \\
&= \frac{D}{2\pi k_B T}[U^{''}(A)U^{''}(B)]^{\frac{1}{2}} e^{\frac{-E_b}{k_BT}} \\
&= \frac{1}{2\pi\gamma}\omega_A \omega_B e^{\frac{-E_b}{k_BT}}
\end{aligned}$$

$$(1.7.15)$$

where E_b= U(B)-U(A) is the barrier height, and ω_A and ω_B are the square roots of the curvature of the potential surface U at A and B . The escape rate falls exponentially with the barrier height. It also decreases when the potential surface at point A and B become flatter, namely the curvatures decrease. On the other hand, the diffusion coefficient D decreases with the increase of the viscosity γ in the solution. All these factor decreases the escape rate r of the particle from the potential well A. Note this only applied to the overdumped region where $E_b >> k_B T$, since we assumed the particle is at quasi-equilibrium and the current density J is close to 0.

1.8 Gas Phase Approaches

In 1999, Bernhard Schlegel et al. investigated the ultrafast photoisomerization of three cyanine dye models ($HN(CH)_n NH^+$, $n = 3,4,5$ correspond to tri-, penta- and heptamethine cyanines, respectively) of different chain lengths using CASSCF quantum chemical calculations.[48] In summary, for all the three models, the photoisomerization processes terminate with torsional motions coupled with the decay in the region of the twisted intramolecular charge-transfer(TlCT)[49] state with an adjacent conical intersection (CI). Excited to S1 state at the planar conformation, the dyes are driven off the Franck–Condon point through a symmetric skeletal stretching mode along a barrierless path. Then a second non-symmetric vibrational mode is dominated by torsional motion about one of the double bonds in the linker that leads the system towards a 90° twisted configuration which is adjacent to the CI. Schlegel et al. found the intersection is not located at the bottom of the S1 surface but near a fully twisted minimum that corresponds to a twisted intramolecular charge-transfer(TlCT) state. After partial equilibration of the TICT state, the skeletal asymmetric stretching and NH wagging modes modified the equilibrium geometry and triggered the nonradiative decay through the CI, leading to the low quantum yield. Inhibition of this trans-cis isomerization by dissolving the dye in viscous medium or chemically rigidizing the linker[50] or by intercalation as found in our work, prevents this decay mechanism, leading to the dramatic fluorescence enhancement.

Gao et al. investigated the photoisomerization of 1,10 -dimethyl-2,20 -pyridocyanine (Me-1122P in Figure 1.8.1) both in gas phase and in methanol.[51] It was found that in the gas phase, conical intersections were near the minima of the S1 state, and the S1 decay follows a barrierless pathway to the global minimum of S1 (minS_1) before relaxing to the S0 state through the CI. The solvent effects were estimated using the polarizable continuum model (PCM)[52]. In methanol as well as other solvents with high polarities such as 1-propanol, ethanol and water, the system would reach a stationary structure (minS_1-trans in Figure 1.8.2) first before the CI, and a significant barrier exists between the stationary structure and minS_1, which results in a much longer lifetime of the excited state.

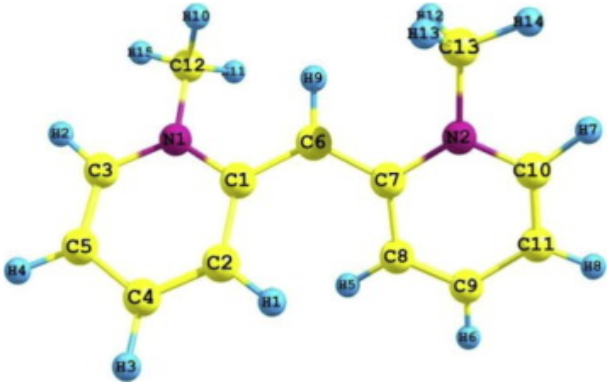

Figure 1.8.1 Me-1122P[51]

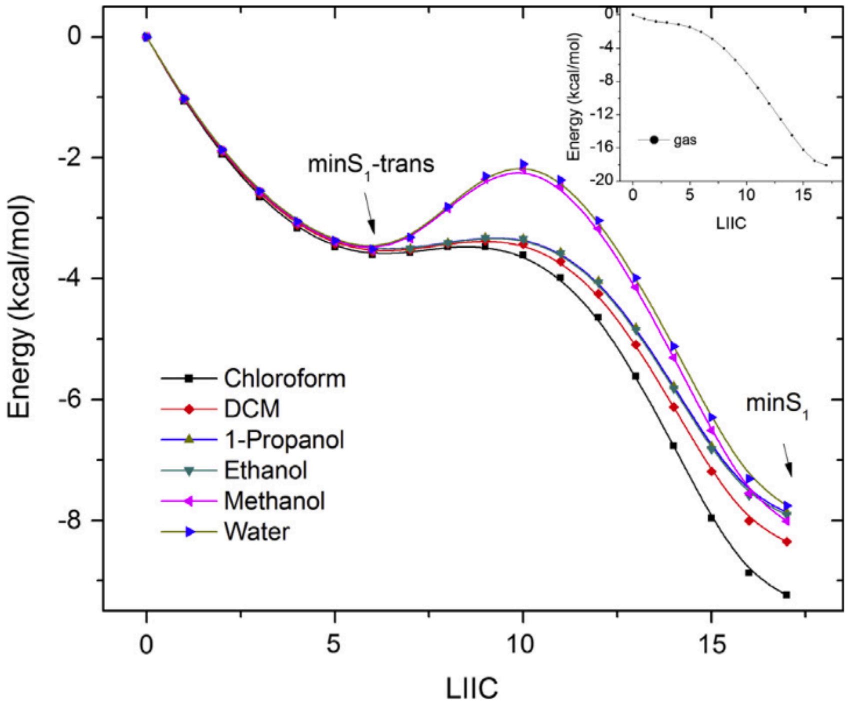

Figure 1.8.2 S1 state energy profiles along the constructed Linearly interpolated internal coordinate (LIIC) pathways .The inset illustrates the energy profile in the gas phase between the Franck-Condon point and minS$_1$. [51]

2 Methods

2.1 Simulations with Modified Amber-Generated Force Fields

2.1.1 Introduction

Oxazole yellow, known as YOPRO (Figure 2.1.1), is a member of the cyanine dye family. It consists of benzoxazole and quinoline rings connected by a linker of two carbon-carbon bonds. YOPRO is almost non-fluorescent in water but its fluorescence is enhanced ~1800 fold after intercalation in double-stranded DNA, providing the basis of DNA concentration assays[1, 8-10]. The rationale for this difference is that in solution, rotational motion of the two rings around the linker permits non-radiative decay to the ground state, while when intercalated, twisting is suppressed, eliminate the non-radiative pathway.

To explore the conformational sampling of YOPRO, we first simulated it in water and when intercalated in double stranded DNA that is stabilized by a basic leucine zipper (bZIP) protein. Both unrestrained and umbrella sampling molecular dynamics were used to obtain the free energy as a function of rotation around the two dihedral angles of the linker. Three different YOPRO force were built by varying the dihedral force constants of the linker bonds, reflecting different assumptions on the linker bonding.

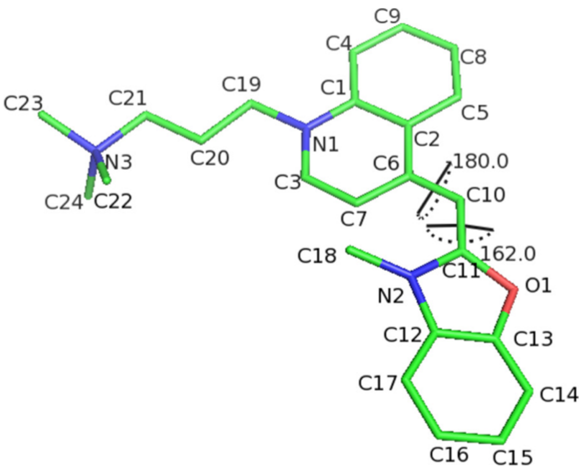

Figure 2.1.1 Structure of YOPRO (without hydrogen atoms) with benzoxazole and quinoline rings connected by a cyanine-based linker. Dihedral C2–C6–C10–C11 (D1) and C6–C10–C11–O1 (D2) are the two dihedrals on linker C11–C10–C6 that define the twist angle of the two ring systems. In this example, D1 = 180° and D2 = 162°.

2.1.2 YOPRO Structure and Internal Force Field Parameters

A structure of YOPRO was obtained from the ChEBI website[53]. Its structure was optimized using Gaussian 03[54] and the Merz-Kollman fitting procedure[55] used to generate partial charges and a mol2 file. Parmchk in Amber[56] was used to generate a frcmod file for tleap. In the frcmod file, we modified the dihedral force constants of the two linker dihedrals to produce three models:

Model 1: The original force field generated by Parmchk has a double bond C6-C10 and a single bond C10-C11. But consideration of the stable resonant structure has a single bond C6-C10 and a double bond C10-C11. Dihedral force field parameters from the Generalized Amber Force Field (GAFF) of 1.0 kcal/mol for each of the nine contributions to single bonds and 6.65 kcal/mol for each of the four contributions to double bond were then used.

Model 2: In the GAFF, the force constant of each of the four torsional energies of a carbon-carbon double bond is set to 26.6/4=6.65 kcal/mol, (the values are half the barriers in accord with the form of the Amber dihedral force field) while that of a single bond is set to 1.4/9= 0.156 kcal/mol. If the numbers of dihedrals involved in a double bond and a single bond are considered, these two force constants are partitioned in the force field by dividing by 4 and 9, respectively. Based on these values, we estimated the force constant for the two YOPRO dihedrals by splitting the difference between single and double bonds; that is, using the geometric average of these partitioned force constants for single and double bonds. The result is close to 1 kcal/mol. This value agrees with the force constant for an sp2 carbon-sp2 carbon bond in the middle of a conjugated system in the GAFF. Thus, we adopt this value for the conjugated cyanine system.

Model 3: Both dihedral force constants were reduced by half to 0.5 kcal/mol.

2.1.3 YOPRO Solvation in Water

The solvation of YOPRO with TIP3P waters was done using tleap.[56] The buffer distance chosen was 8 Å. To achieve electroneutrality of the system, 2 chloride counterions were added. Thus, the YOPRO was surrounded by 751 water molecules and 2 Cl^- ions (which made the total size of the system 2312 atoms). A restrained energy minimization was performed using the SANDER module of AMBER12. For both the minimizations and the subsequent MD runs, the long-range Coulombic interactions were handled by the particle mesh Ewald method. The minimizations were performed in 2 stages. In both stages 4000 iterations of steepest descent were performed followed by 4000 iterations of conjugate gradient algorithm. A force constant of 10 kcal/mol/rad^2 was applied to the restrained atoms. During the first minimization, the heavy atoms of YOPRO were restrained, which allowed for the optimization of the hydrogens of the waters added by tleap. While in the second one, all atoms in the system are minimized.

After the restrained minimizations, a 200 ps heating run was done to 300 K. The heat run used the exact same restraints that were used in the minimizations except that a force constant of 100 kcal/mol/rad^2 was applied to the restrained atoms. SHAKE[57] was used in the heating and subsequent production runs for bonds involving hydrogen atoms allowing for a 2-fs time step. A Langevin thermostat[56] was used to maintain the temperature at 300 K. The heated structures were used as the reference structures for the next 800 ps equilibration run, which allowed the solvent environment to reach their proper densities. At the end of the equilibration run, the density of the YOPRO system was 0.9922 g/cc. The restart files from those equilibration runs were then used as the inputs for all of the subsequent production runs. Then MD runs were performed at 300 K using constant NPT conditions (isothermal-isobaric) for the three models as detailed in Chapter 3. The reference pressure was set equal to 1 bar, and the Berendsen barostat[58] used with a pressure-coupling constant of 0.1 psec. The temperature was maintained at 300 K using a Langevin thermostat with a collision frequency of 0.2 ps^{-1}.

2.1.4 bZIP-DNA Solvation in Water

The crystal structure of GCN4 in the presence of DNA (PDB accession code

1YSA) was used[59] to initiate the simulations with bZIP. It was solvated with 22526 TIP3P waters with an 8 Å buffer distance. 25 Na$^+$ counterions were added to achieve electroneutrality. The minimization and heating runs were performed using the same restraints that were used for the "YOPRO in water" system. The heating run lasted 200 ps and was followed by a 400 ps equilibration run. At the end of the equilibration run, the density of the bZIP system was 1.0070 g/cc. The restart file from those equilibration runs were then used as the inputs for the subsequent production MD runs. The 20 ns MD run was performed at 300 K using constant NPT conditions with reference pressure 1 Bar and pressure-coupling constant of 0.1 psec using a Berendsen barostat.[58] The temperature was maintained at 300 K using a Langevin thermostat with a collision frequency 2 ps^{-1}.

2.1.5 bZIP-DNA-YOPRO Intercalation in Water

We used PyMOL[60] to intercalate YOPRO into the space between the 2 base pairs C4-G38 and T5-A37 in the bZIP structure to obtain a coordinate file for the bZIP-DNA-YOPRO complex. The two oxygen atoms on end bases 1 and 21 were deleted to make the structure compatible with tleap. The bZIP-DNA-YOPRO complex was solvated with 20524 TIP3P waters with a buffer distance of 8 Å. 21 Na$^+$ counterions were added for electroneutrality. The minimization and heating run were performed using the same restraints that were used for the YOPRO in water system. The heating run lasted 200 ps and was followed by the equilibration run. The heated structures were used as the reference structures for the next 200 ps equilibration run, which let the protein maintain its original structure while the solvent environments were allowed to reach their proper densities. At the end of the equilibration run, the density of the bZIP-DNA-YOPRO complex was 1.011g/cc. The restart file from those equilibration runs were then used as the inputs for the subsequent production MD runs. After equilibration, different starting dihedral angles were used, as detailed in the following sections.

2.1.6 Umbrella Sampling of YOPRO Model 1 and Model 2 in Water

For the model 1, the umbrella sampling of the dihedral C2-C6-C10-C11 referred to as D1 started from -110 degrees using the restart file from the free simulation. The windows were spaced 20 degrees apart from -170 to 170 degrees. A force constant of 20 kcal/mol/rad^2 was used for the windows from -130 to -10 degrees while the force constant of all other windows was 40 kcal/mol/rad^2. Each window was started with a 200 ps energy minimization consisting of 500 iterations of steepest descent and 1500 iterations of conjugate gradient algorithm. The next step was 200 ps equilibration. The same conditions as for unrestrained runs were used. A Berendsen thermostat was used to heat the system from 0 to 300 K under constant pressure (1 bar) with a pressure relaxation time of 5 ps. Each window production run was 1 ns.

For the model 2, two dimensional umbrella sampling was carried on D1 and D2 (dihedral C6-C10-C11-O1) starting from D1=–120 and D2=0 degrees. The restart file from the free simulation was used. For D1, the "regular" windows were spaced 20 degrees apart from –180 to 180 degrees with the addition of several windows at the extremes for better overlapping (D1=±19 and ±10 degrees). D2 was limited to the interval [-40, 40] with 5 windows spaced 20 degrees apart because the free simulation results indicated the probability density distribution concentrated only in that range. The force constant for D2 was always 20 kcal/mol/rad^2, while that of D1 was also 20 kcal/mol/rad^2 in all window except those six where the force constant of D1 was set to 60 kcal/mol/rad^2: (D1=−10, D2=0, ±20, ±40) and (D1=10, D2=−40). All constraints used except the force constants were the same as those in Model 1. Each window was simulated for 10 ns after 1ns equilibration. All the windows overlap with their neighboring ones very well (Figure 2.1.2).

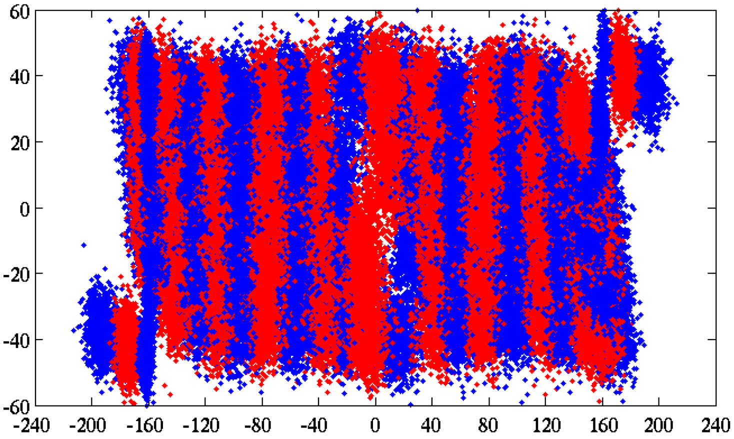

Figure 2.1.2 This density distribution shows that these 105 windows mentioned above cover the whole region of D1∈ [-180,180] and D2∈ [-40,40] for the model 2.

2.1.7 Unrestrained Intercalation Simulation of Model 1, 2 and 3

For the model 1, a 10ns simulation of D1 was started from −165.0 degrees. For both model 2 and 3, D1 was started from −152.4 degrees and D2 was started from −4.6 degrees, after equilibration. Two successive trajectories of 10 and 13 ns were run for Model 2 and one trajectory of 13 ns was run for Model 3.

2.1.8 Umbrella Sampling Simulation of Models 2 and 3 Intercalated

Only D1 was scanned, while D2 was constrained around 0 degrees (for model 2) and 20 (for model 3). The restart file from the free simulation was used for start up. The sampling was started from D1=–160 degrees (for model 2) and –140 degrees (for model 3) and then extended to both positive and negative directions to cover the whole range [–340,10]. After unbiasing by WHAM, this dihedral was converted to [-180,180] for plotting. The force constant of D2 was always 20 kcal/mol/rad^2. The force constants of D1 are given in Table 2.1.1 and 2.1.2:

D1	Force constant (kcal/mol/rad2)
-240 to -340, every 10 degrees	60
-170 to -230,every 10 degrees	40
-160 to -120,every 10 degrees	20
-110 to -50, every 10 degrees	40
-40 to -20, every 10 degrees	60
-10 to 0, every 5 degrees	80
1 degree	100
3 degrees	160
5 degrees	200
8 degrees	200
10 degrees	100

Table 2.1.1 Window data for Umbrella sampling of Model 2 intercalated.

D1	Force constant (kcal/mol/rad^2)
15 degrees	160
10 degrees	160
5 degrees	140
0 degree	120
-2 degrees	120
-5 degrees	120

-7 degrees	120
-10 degrees	100
-12 degrees	100
-15 degrees	60
-19 degrees	80
-20 to -30, every 10 degrees	60
-40 to -110, every 10 degrees	40
-120 to -160, every 10 degrees	20
-170 to -230, every 10 degrees	40
-240 to -280, every 10 degrees	60
-290 to -340, every 10 degrees	80

Table 2.1.2 Window data for Umbrella sampling of Model 3 intercalated

2.2 Simulations with Fitted Force Fields by the Implicitly Polarized Charge Method

2.2.1 Partial Charge Set Fitting for Water-solvated YOPRO

The Implicitly Polarized Charge Method(IPolQ)[11] described in Section 1.3 was used to fit for the partial charge set for water-solvated YOPRO.

After solvating YOPRO in 2238 SPC water molecules, we froze the YOPRO and carried out 10 ns conformational sampling of the solvent at 450K to collect a set of equilibrated water configurations.

Next, the MD simulation of the solute YOPRO were performed from each equilibrated water configuration with the solvent fixed to obtain a set of time averaged solute configurations. The fixed locations of the water solvent served to create a field of point charges surrounding the solute.

Once the MD simulations were finished, we used orca[34] to calculate the single point energy and electrostatic potential of each solute configuration in vacuum for charge fitting at B3LYP/cc-pvTZ level, and took the solvent charge density as a perturbation for the electrostatic potentials in water. The final IPolq charge set was derived from the perturbed electrostatic potential set.

The IPolQ charges set was used as the input for the next iteration by replacing the original guess. Once the fitted charge set is converged, the charge fitting process is finished. The IPolQ charge then goes to the force constant fitting section.

2.2.2 Dihedral Force Constant Fitting

The force constant fitting method described in Section 1.5 was used to fit for the force constant parameter of D1 and D2 (figure 2.1.1). To start with, 36×36 conformations of YOPRO were generated by applying NMR constraints on the two dihedrals, in vacuum.

The following are some fitted force field parameters (a_n and b_n values):

$D2^{a, b}$	a2	a4	a6	a8	b2	b4	b6	b8
-175	8.595	4.010	0.645	1.836	174.181	-20.620	101.500	155.290
-125	10.940	2.592	2.462	0.867	165.630	-113.100	91.120	-145.600
-75	12.715	2.286	1.456	1.829	202.610	55.780	-107.900	-120.900
-25	12.060	2.026	1.684	0.917	194.400	7.009	240.300	-70.410
25	11.420	2.217	1.458	2.348	171.478	-37.760	125.800	94.930
75	11.440	2.221	0.502	0.940	165.770	-45.790	125.700	-1.898
125	11.470	2.978	2.003	0.246	188.932	143.040	-77.370	-106.500
175	8.920	1.405	2.334	3.090	188.786	22.430	162.860	-105.300

Table 2.2.1 S0 YOPRO a and b values of D1 surface as functions of D2

$D1^{a, b}$	a2	a4	a6	a8	b2	b4	b6	b8
-175	9.775	3.437	1.02	1.091	174.767	-0.924	225.55	192.52
-125	10.515	1.615	1.173	1.704	168.71	3.323	105.8	147.17
-75	12.305	2.829	1.633	1.595	216.69	69.53	247.71	-121.3
-25	8.91	1.683	0.457	0.431	187.063	-28.27	130.61	166.36
25	11.15	1.53	0.936	1.871	167.91	0.261	81.66	137.3
75	11.625	2.368	1.279	1.564	150.78	-32.03	-211.5	244.04
125	11.235	2.846	1.274	0.638	191.13	78.67	-76.14	244.86
175	8.175	3.289	0.783	1.872	184.417	14.44	181.946	208.43

Table 2.2.2 S0 YOPRO a and b values of D2 surface as functions D1

$D2^{a, b}$	a2	a4	a6	a8	b2	b4	b6	b8

-175	6.105	3.742	3.375	1.806	8.553	195.350	-16.860	-14.480
-125	10.145	7.520	2.179	5.580	-38.470	145.020	-113.800	206.900
-75	2.409	2.747	3.171	3.212	-90.120	132.980	-86.530	-97.600
-25	6.815	5.790	4.004	3.725	-27.000	178.759	0.349	3.228
25	7.130	3.825	4.503	1.315	-7.339	166.330	-23.730	-24.470
75	2.326	1.223	1.960	2.091	202.530	179.736	48.600	41.410
125	9.795	9.875	2.198	5.295	37.870	210.960	86.170	166.440
175	5.195	3.522	3.143	4.763	15.890	177.134	-5.574	-4.631

Table 2.2.3 S1 YOPRO a and b values of D1 surface as functions of D2

The torsional force constants of D1 are minimized at D2= ±75, which corresponds to high energy regions on the quantum energy surface of S1 YOPRO.

D2^{a,b}	a2	a4	a6	a8	b2	b4	b6	b8
-175	11.105	2.344	0.914	0.565	183.751	-7.855	-95.060	92.930
-125	13.855	3.599	2.018	0.195	188.427	63.950	-55.720	589.300
-75	12.830	8.380	4.466	1.256	146.220	-22.980	208.160	164.430
-25	10.140	0.889	2.142	1.866	177.814	-114.000	-9.114	31.450
25	11.160	0.985	0.905	0.564	174.326	247.670	84.830	6.343
75	19.335	12.31	5.855	1.534	209.330	18.080	170.567	198.240
125	14.565	2.001	1.653	0.452	175.116	-48.170	61.610	9.348
175	10.055	3.188	0.230	0.629	180.027	29.220	-21.640	-93.940

Table 2.2.4 S1 YOPRO a and b values of D2 surface as functions D1

The torsional force constants of D2 are maximized at D1= ±75, which corresponds to low energy regions on the quantum energy surface of S1 YOPRO.

The final fitted result shows that in the S0 state, the two linker bonds are more or less in the same bond order, as expected for the ground state of YOPRO, while in the S1 state, D1 (see Figure 2.1.1) turns to be more like a pure single bond and D2 more like a pure double bond.

2.2.3 Umbrella Sampling of YOPRO in Water on the S0 and S1 Surfaces

Two dimensional umbrella sampling on the S0 surface was carried out on dihedrals D1 (C2-C6-C10-C11) and D2 (C6-C10-C11-O1). Note that the dihedral potentials that we developed, coupling D1 and D2, are well-suited to the umbrella sampling method whereby restricted window ranges for these dihedrals are used. Otherwise, in a free simulation, the specific dihedral-dependent force field would have to be introduced when the dihedrals' would reach different regions of their angles. That would require monitoring the dihedral values and stopping and restarting the simulation for the new force field values.

The windows were started from the trans-cis conformation (180, 0) which is around the global minimum on the quantum chemical energy surface with the corresponding fitted internal force constant sets. The starting window was initialized with a minimization followed by a 5ns equilibration at 300 K using constant NPT (isothermal-isobaric) conditions, and the restart file was passed to start the equilibration of its neighboring windows. Each adjacent window was spaced by 10 degrees along D1 or D2. Finally, the sampling range expanded to the full space ($360^0 \times 360^0$) covered by 36×36 windows. Once all the windows were equilibrated, their production runs were started at the same time. All minimization, equilibration and production runs were carried out with a restraint force constant of 10 kcal/mol/rad^2 for both D1 and D2.

All the umbrella samplings were unbiased with the WHAM methodology[61, 62] to obtain the two dimensional free energy surfaces along the D1 and D2 dihedral angles.

For the S1 surface, two ways of generating the window data were used. The first way was started at the same places as in S0 sampling. The exact same protocol was used, except that the internal force fields of YOPRO for each window were those for the S1 state.

In the second way, the umbrella sampling was started at all four minima: (0,0), (0,180), (180,0) and (180,180) on the "subtracted" S0 surface of YOPRO for the force constant parameter fitting as described in Section 1.5 with the corresponding fitted internal force constant sets for the S1 state. The same protocols were used as in the first method, except this time we sampled four regions covered by 15×15 windows with the four starting windows in the center.

This protocol is used since the minima on the S0 surface are the Frank-Condon transition points as discussed in Section 1.6. Later on this protocol is used for the S0 and S1 intercalations as discussed in the next section because during the full space sampling of intercalated YOPRO, deintercalation always happens which must be avoided because YOPRO never go back to the "box" after that.

2.2.4 Umbrella Sampling Simulation of Intercalated YOPRO Surfaces

The same protocols were used to sample intercalated YOPRO in the S0 and S1 states as for the second sampling method of S1 YOPRO in water outlined in Section 2.2. When the production runs were finished, VMD[63] was used to view the trajectories of the windows to exclude those where YOPRO de-intercalated.

3 Results

3.1 bZIP-DNA in Water

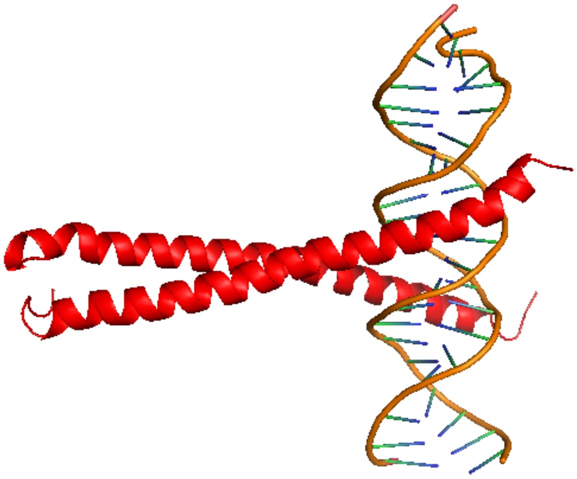

Figure 3.1.1 The bZIP crystal structure[59] showing the two monomers composed of leucine zipper and basic domains. The dsDNA that is perpendicular to the protein dimer is held in place by the basic domains.

The bZIP-dsDNA complex is comprised of 40 paired bases and two leucine zipper monomers each with 57 residues. The crystal structure of GCN4 in the presence of DNA (PDB accession code 1YSA)[59] was used to initiate the simulations (see Chapter 2). Figure 3.1.1 displays the two alpha-helical monomers and the DNA that is bound essentially perpendicular to dimerized bZIP.

The bZIP-DNA was first simulated without YOPRO intercalated for 20 ns. Root mean square fluctuations (RMSFs) were traced for DNA (nucleic acid residues) and bZIP (protein residues) separately, as shown in Figure 3.1.2. For the double stranded dsDNA we ignored the bases at the beginnings of the two strands so the RMSFs of DNA start from base 2 and end at base 40, with one strand labeled as 2-20 and the other 22-40. The DNA is quite stable, with greater fluctuations at its extremities where it is not as constrained by interactions with bZIP.

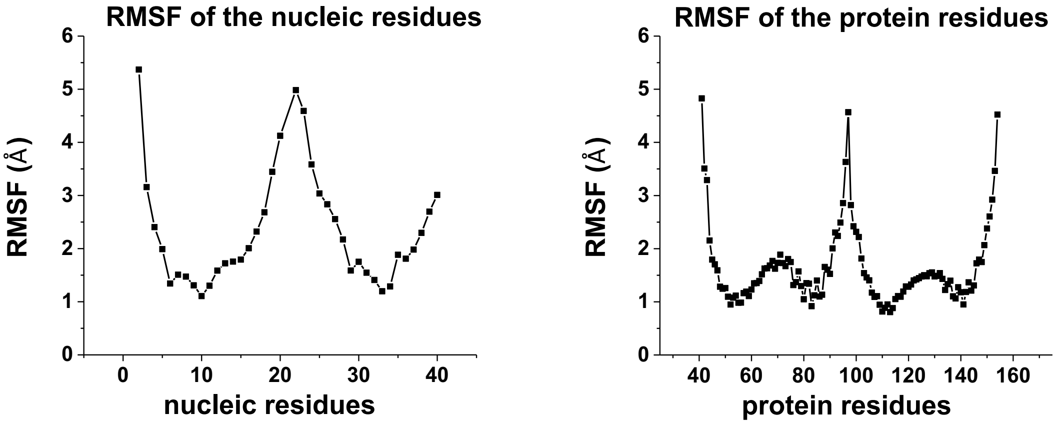

Figure 3.1.2 RMSFs of the DNA bases and protein residues for bZIP-DNA. The dsDNA strands are numbered 2-20 and 22-40. The protein monomers are numbered 41-97 and 98-154. The bases and residues are quite stable with the protein dimer maintained and base pairing maintained with greater fluctuations at the ends of both monomers and both DNA strands.

In order to simulate bZIP with intercalated YOPRO space must be provided between paired bases. The procedure and results for doing so are now presented.

3.2 Preparing Space in bZIP-DNA for YOPRO intercalation

Based on experiments intercalating YOYO, which consists of two YOPRO moieties connected by a bridge region, sites for YOPRO intercalation were found. The rings of each YOPRO predominantly intercalated in a double stranded, palindromic dsDNA consisting of 16 or 24 residues. The YOYO rings intercalate preferentially to d1(5'-CTAG-3') : d2(3'-GATC-5') binding sites. [64] Therefore, we assume that YOPRO intercalates into d1(CT):d2(GA) and d1(AG):d2(TC) binding sites. We picked one d1(5'-C4T5-3'):d2(3'-G38A37-5') and used constraint MD to gradually separate these neighbor base pairs to provide intercalation space for YOPRO. The expansion protocol is detailed in the Methods section.

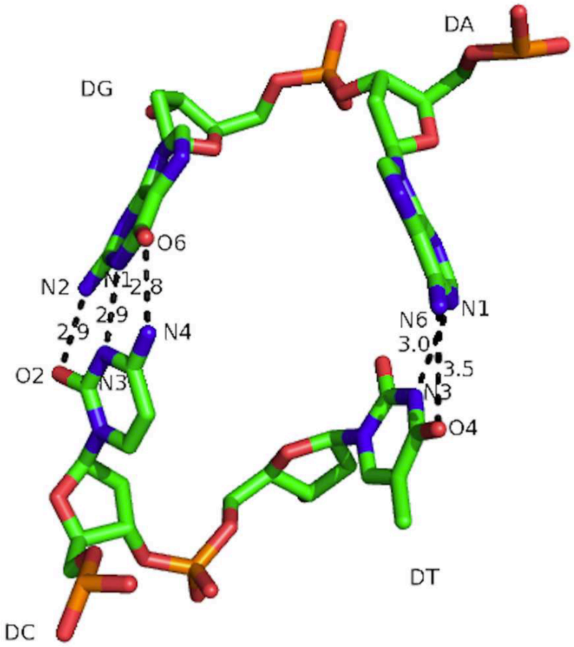

Figure 3.2.1 The hydrogen bonded base pairs, G38-C4 and A37-T5 separated for YOPRO intercalation, with the three GC and two AT hydrogen bond distances at the end of the separation indicated.

Figure 3.2.1 displays the endpoint of the separation simulation and Table 3.2.1 provides key distances. The figure shows that the five hydrogen bonds between the

GC and AT bases are preserved. The distances that measure the expansion of the space between the neighbor base pairs increase from about 3.5 to 9.4 Å. This about 6 Å increase in separation is required to intercalate YOPRO.

Atom pairs	C4C2-G38C4	G38C4-A37C2	C4C2-A37C2	T5C5-A37C2	C4C2-T5C5	T5C5-G38C4
Starting Distance (Å)	6.08	3.59	7.06	5.49	3.37	6.44
Ending Distance (Å)	6.08	9.39	11.19	5.49	9.37	10.9

Table 3.2.1 Box expanded atom pair distances.

Plots of the hydrogen bond distances (Figure 3.2.2) show that all five are well maintained throughout the expansion.

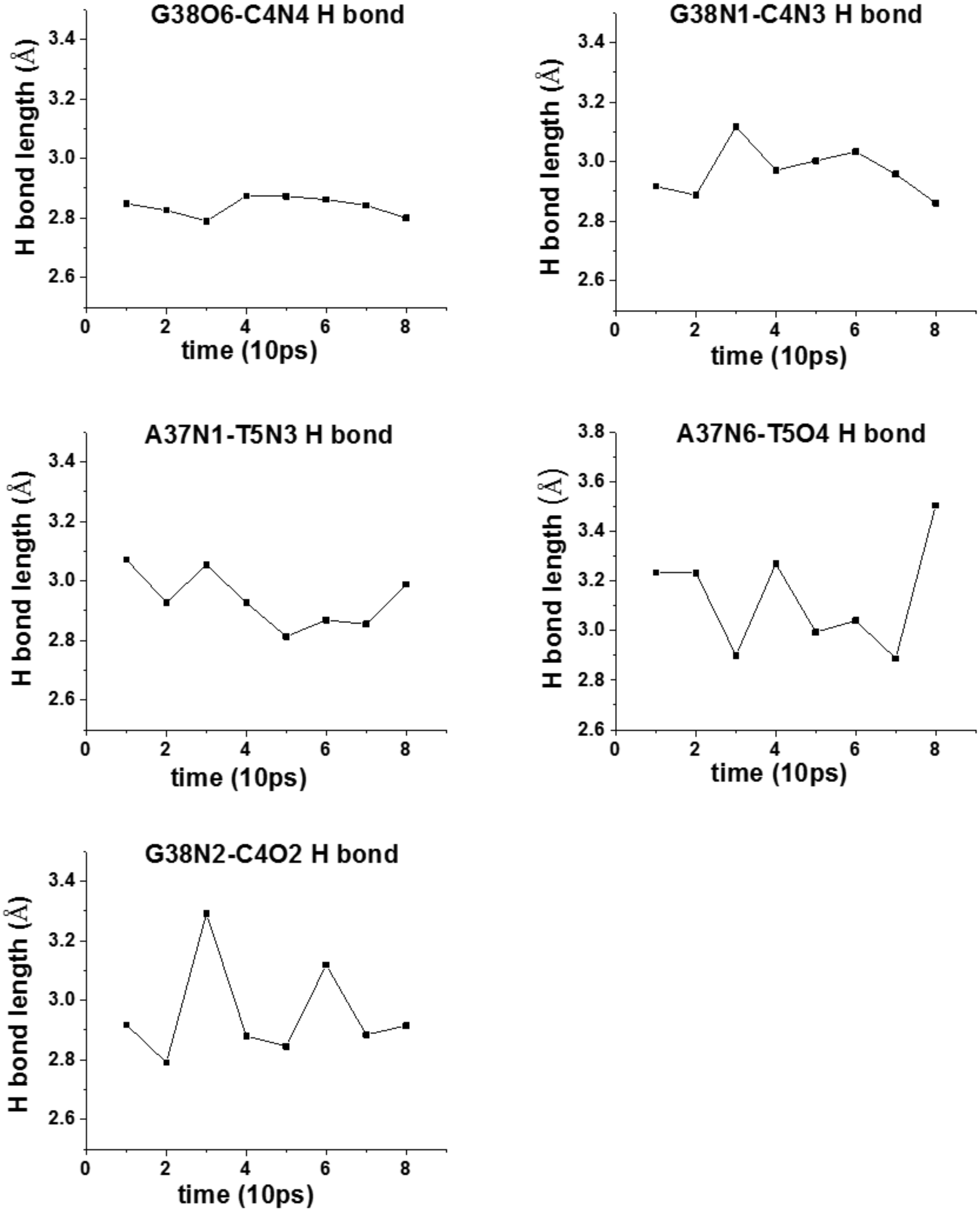

Figure 3.2.2 During the expansion to provide room for YOPRO intercalation, the base pairing hydrogen bonds are well maintained.

3.3 YOPRO Models 1,2 and 3

3.3.1 YOPRO Model 1 Intercalation in bZIP-DNA

The structure of YOPRO is given in Figure 2.1.1.1. As commonly drawn[64, 65], the linker between the benzoxazole and the quinoline ring systems consists of a single bond (C11-C10) and a double bond (C10-C6). Therefore, the dihedral defined by C2-C6-C10-C11 is sufficient to represent the angle between the two rings systems. This YOPRO will be referred to as YOPRO Model 1. Standard internal force field parameters for these single and double bonds were set by the Generalized Amber Force Field (GAFF), as discussed in Chapter 2.

YOPRO was manually docked into the 'box' in Figure 3.2.1, and 10 ns MD of this bZIP-DNA-YOPRO complex initiated. Figure 3.3.1 shows the intercalated YOPRO along with some distances that are used to characterize its bounding box.

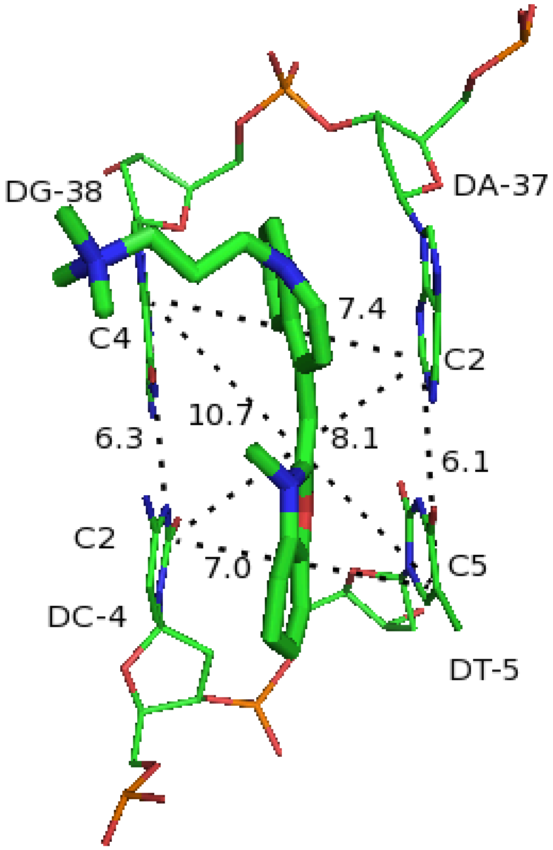

Figure 3.3.1 A snapshot from the 10 ns MD trajectory of YOPRO intercalated into the 'box' formed from the separating the base pairs. Six distances are used to characterize the stability of the box: A37C2-C4C2, T5C5-C4C2, G38C4-T5C5, G38C4-C4C2, A37C2-T5C5, and A37C2-G38C4.

The five GC and AT hydrogen bonds and the box dimensions were well maintained during the simulation, as summarized in Tables 3.3.1-3.3.2.

H bonds	Average bond lengths (Å)	Standard deviations (Å)
A37N1-T5N3	2.944	0.1082
A37N6-T5O4	3.017	0.1921
G38N1-C4N3	2.961	0.1065
G38N2-C4O2	2.896	0.1309
G38O6-C4N4	2.969	0.1965

Table 3.3.1 Hydrogen bond averages and standard deviations of the YOPRO-intercalated bases over a 10 ns simulation.

Atom pairs	Average distances (Å)	Standard deviations (Å)
A37C2-C4C2	8.739	0.5273
A37C2-G38C4	7.754	0.5518
A37C2-T5C5	5.992	0.1656
G38C4-C4C2	6.268	0.1157
G38C4-T5C5	10.407	0.5296
T5C5-C4C2	7.408	0.5799

Table 3.3.2 Averages and standard deviations of the YOPRO-intercalated box dimensions over a 10 ns simulation.

In addition, the location of YOPRO relative to the box was parameterized by some distances between atoms belonging to YOPRO and the bases, as displayed in Figure 3.3.2.

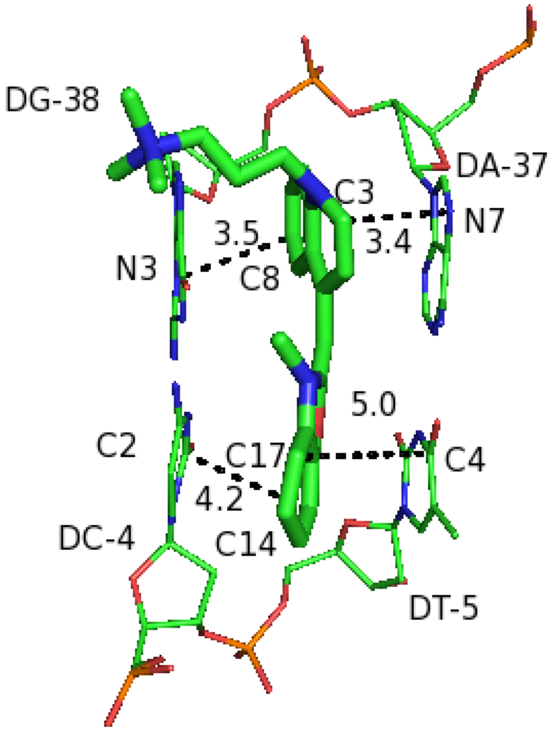

Figure 3.3.2 Base-to-YOPRO distances used to characterize the intercalation extent, for a snapshot from the 10 ns MD trajectory.

Monitoring these distances over the MD trajectory (see Table 3.3.3) shows that YOPRO always remains inside the box in the sense that both rings are kept within the enclosure of the four bases.

Atom pairs	Ave distances (Å)	Standard deviations (Å)
A37N7-YOPRO C3	4.3574	0.9552
C4C2-YOPRO C14	5.5450	1.9467
G38N3-YOPRO C8	3.8205	0.5871
T5C4-YOPRO C17	5.7346	1.4060

Table 3.3.3 Averages and standard deviations of distances between G, C, A and T and indicated YOPRO atoms over the MD trajectory.

The propanamine "tail" does stick out of the base enclosure; however, what is key is that the two ring systems remain well intercalated during the MD trajectory.

3.3.2 YOPRO Model 1 in Water

Using the model 1 force field (see Section 2.1), two simulations of YOPRO Model 1 were carried out in water. First, a 10 ns simulation started from a configuration with the single bond dihedral D1 (C2-C6-C10-C11) at angle −170.5 degrees, obtained from the intercalated structure. This unrestrained simulation covered a small dihedral range. A histogram of the dihedral probability shows that the dihedral fluctuated around −120 degrees. This is far from a planar conformation of the two ring systems. However, even in water, MD may not indicate the global free energy minimum of a system by not being able to surmount barriers in the complex potential energy surface. Therefore, we carried out standard umbrella sampling[66, 67] to obtain the probability of observing this dihedral angle. Nineteen windows, each simulated for 1 ns, were used to cover the range (−180,180), as detailed in the Section 2.1. The umbrella constraints were unbiased with WHAM[27, 61, 62]. Figure 3.3.3 shows that the dihedral angle is indeed restrained to about −115 degrees while the unrestrained simulation dihedral probability agrees well over its limited range with that obtained from the umbrella sampling.

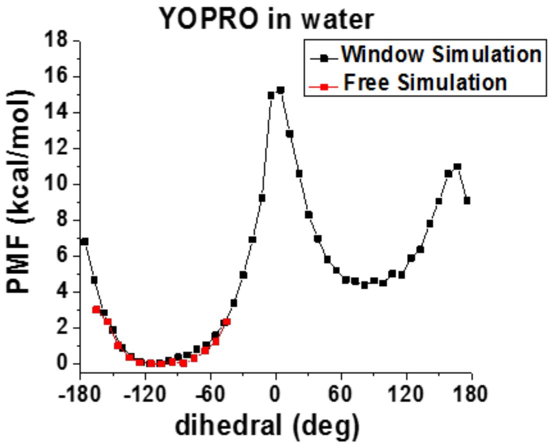

Figure 3.3.3 PMF of the dihedral D1(C2-C6-C10-C11). The red line is the result of the unrestrained simulation while the black line is obtained from umbrella sampling.

The second minimum at around 90 degree is about 4 kcal/mol higher than the global minimum, indicating that it has a population, $P \sim 0.0007$, relative to the global minimum. Thus, in water the YOPRO the single bond C2-C6-C10-C11 dihedral is quite well restricted to sampling around −120 degrees.

3.3.3 YOPRO Model 1 Intercalated

To see if the intercalated YOPRO model 1 would lead to a more planar configuration of the ring systems, we simulated the intercalated bZIP-DNA-YOPRO complex for 10 ns starting from the geometry at the end of the docking intercalation run in Section 3.3.1, when D1 was −165.0 degrees. A histogram of D1 is displayed in Figure 3.3.4.

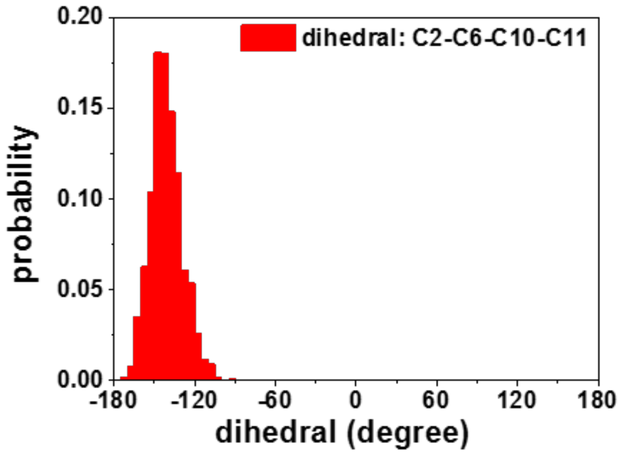

Figure 3.3.4 PMF of the dihedral D1(C2-C6-C10-C11). The red line is the result of the unrestrained simulation while the black line is obtained from umbrella sampling.

The unrestrained simulation of intercalated YOPRO shows that this dihedral fluctuates in a range around −145 degrees, which is not very different from that obtained with YOPRO in water (see Figure 3.3.4). Thus, the differentiation between YOPRO in water and intercalated is not evident here. Of course, it is possible that if a different initial D1 angle were used, YOPRO could fluctuate around some other minimum angle. However, rather than pursue an umbrella sampling simulation to obtain a PMF around this dihedral, a more realistic model for the YOPRO linker will be now introduced in the next section.

With regard to the overall structure of the bZIP-DNA-YOPRO complex, Figure 3.3.5 shows the RMSFs of the bases and the residues for the complex, along with those for bZIP-DNA. The RMSFs are all quite similar. Thus, both in terms of overall structure and the specifics of the YOPRO intercalation, the bZIP-DNA-YOPRO complex is stable and the YOPRO ring systems are properly intercalated within the four bases.

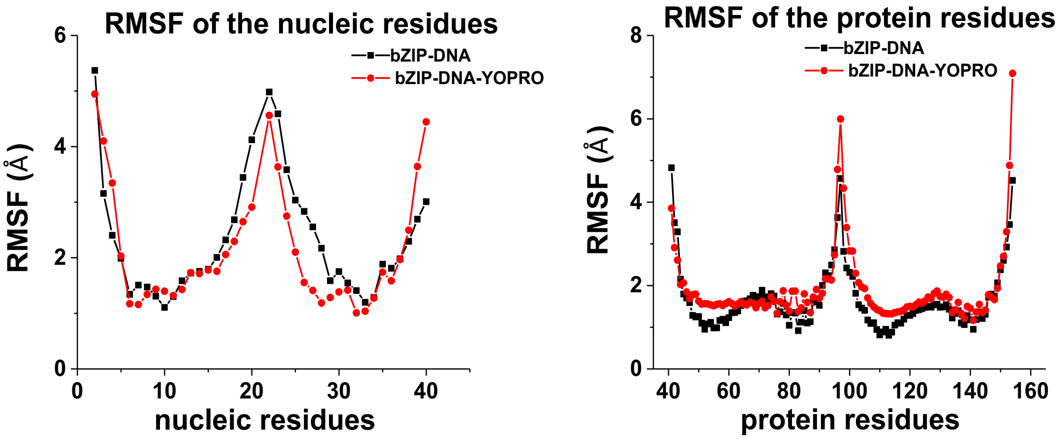

Figure 3.3.5 Model 1 RMSFs of the DNA bases and bZIP residues compared for bZIP-DNA and bZIP-DNA with intercalated YOPRO over their MD trajectories.

3.3.4 Revision of the YOPRO Force Field, Model 2

The small difference between the YOPRO single bond dihedral angle in both water and when intercalated brought us to the idea of reconsidering the accuracy of the force field for the atoms forming the linker between the ring systems. Indeed, the quantum chemical optimization (see Section 2.1) showed that the two bonds forming the linker are essentially the same length, 1.4 Å, and, therefore, should be of the same bond order, as appropriate to a more-or-less symmetrical cyanine dye. Cyanine dyes with the structure Aryl=N$^+$=CH[CH=CH]$_n$-N=Aryl are generally considered as delocalized systems.[68, 69] That seems to be the case here. Thus, the two linker bonds are considered as delocalized "bond-and-a-half" bonds and modifications to the generalized amber force field (GAFF) were made. The result is that the torsional angle force constants for the two dihedrals D1 and D2 were set to values that interpolate single and double bonds (see Section 2.1 for details). This model will be referred to as YOPRO Model 2.

3.3.5 YOPRO Model 2 in Water

We carried out a 270 ns simulation of YOPRO in water with this revised force field. Now the dihedral angles around the both bonds in the YOPRO linker were monitored. D1 is the original dihedral, while the new one, D2, is defined as C6-C10-C11-O1 (see Figure 2.1). Figure 3.3.6 shows that D2 fluctuates around zero degrees, while D1 now has four positions of high probability: ±70 and ±140 degrees, which is quite different from the previous result shown in Figure 3.3.6.

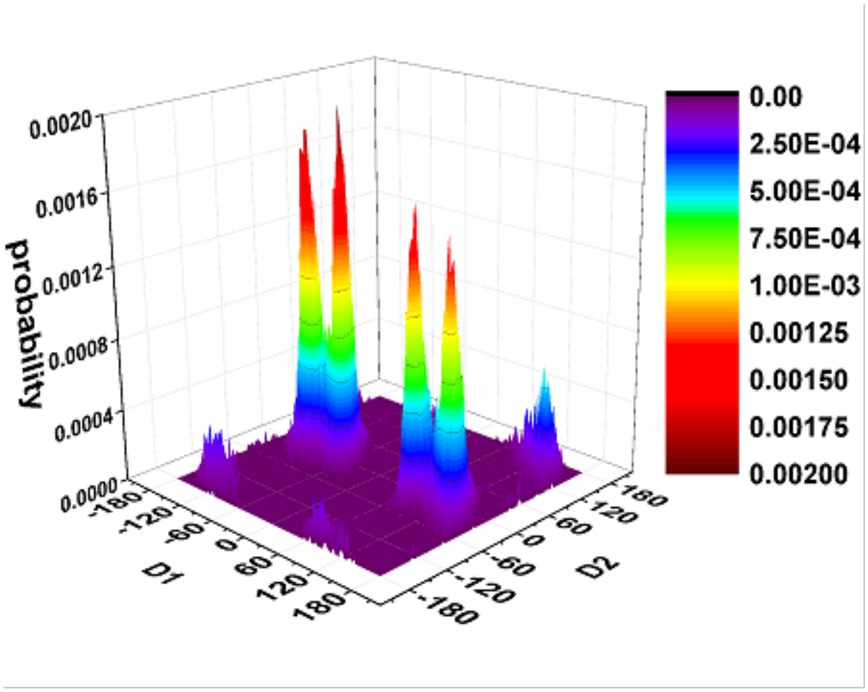

Figure 3.3.6 The probability density for the two dihedrals for the unrestrained YOPRO Model 2 in water.

We also carried out two dimensional umbrella sampling to support this result, as detailed in the Section 2.5. A total of 105 windows were used to sample D1 from –180 to 180 degrees and D2 from −40 to 40 degrees, to cover the high probability ranges of the unrestrained simulation. The probability are displayed in Figure 3.3.7.

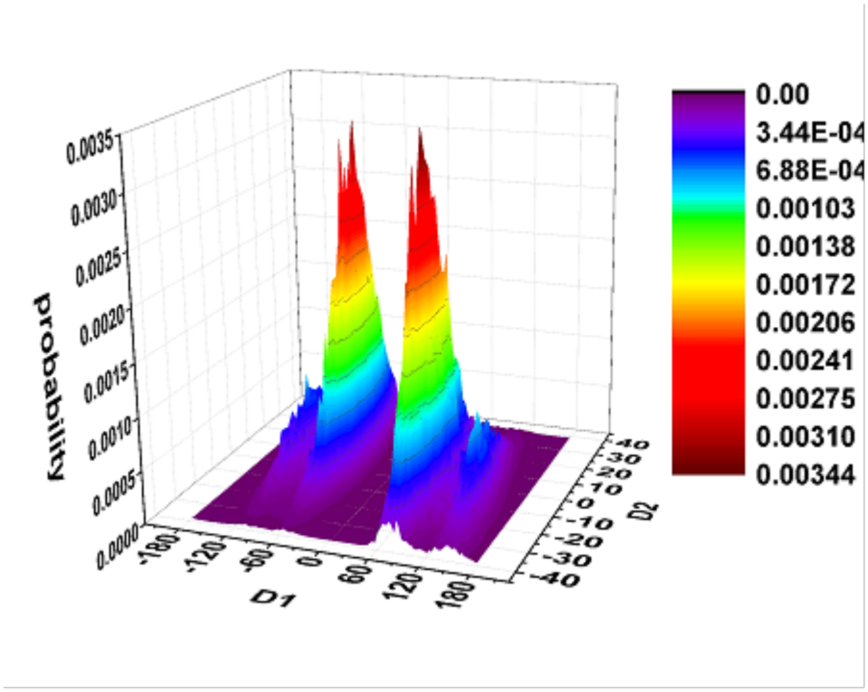

Figure 3.3.7 The probability density (top) for the two dihedrals from the 2D umbrella sampling simulation.

This result is in good agreement with that obtained from the unrestrained simulation in the sense that D1 peaks are at ±65 and ±130 degrees, the same position as those in the free simulation.

To investigate why the probability distributes like this, we used PYMOL[60] to constrain D2 at 0 and rotated the benzoxazole ring around the bond between C6 and C10 and measured the distances between different atom pairs every 20 degrees. As is evident in Figure 2.1, the C7-O1 and C5-O1 are two atom pairs for which their distances could sample smaller than van der Waals contact. Assuming that these distances might be the dominant factors in the PMF, we evaluated the sum of the Lennard-Jones potentials of these two atom pairs at different D1 values. The results show that when D1 approaches zero degrees, both C7-O1 and C5-O1 distances become very short, leading to Lennard-Jones potential energies that are much larger than thermal. This clash explains much of the structure of Figure 3.3.6. There is more structure in the probability surface of the umbrella sampling. However, the slight enhanced probability around ±65 degrees corresponds to only about a 1 kcal/mol barrier in the PMF, which is on the thermal energy scale.

3.3.6 YOPRO Model 2 Intercalated

We then carried out a 10 ns simulation on intercalated YOPRO with this modified force field. Table 3.3.4 shows that the five intercalating base-base hydrogen bonds are well maintained and Table 3.3.5 shows that the atom pair distances of the box that confines YOPRO are stable.

H bonds	Average bond lengths (Å)	Standard deviations (Å)
G38N1-C4N3	2.952	0.08896
G38O6-C4N4	2.926	0.1410
A37N6-T5O4	2.996	0.1968
A37N1-T5N3	2.956	0.1086
G38N2-C4O2	2.866	0.1201

Table 3.3.4 The intercalating hydrogen bond average lengths and standard deviations over the 10 ns simulation.

Atom pairs	Average distances (Å)	Standard deviations (Å)
A37C2-C4C2	8.535	0.4529
A37C2-G38C4	7.449	0.4120
A37C2-T5C5	6.007	0.1595
T5C5-C4C2	7.166	0.6327
G38C4-C4C2	6.286	0.1082
G38C5-T5C4	10.137	0.7354

Table 3.3.5 The box atom pair average lengths and standard deviations over the 10 ns simulation.

The distance monitors for the YOPRO ring systems remaining between the confining bases displayed in Table 3.3.6also show that the intercalation is stable.

Atoms pairs	Average distances(Å)	Standard deviation (Å)
C4C2-YOPRO C14	4.356	0.5716
T5C4-YOPRO C17	4.283	0.4600
A37N7-YOPRO C3	3.966	0.4196
G38N3-YOPRO C8	4.900	0.9586

Table 3.3.6 The atom pairs average lengths and standard deviations (see Figure 7) for the distances used to characterize the intercalation extent over the 10 ns simulation.

Figure 3.3.8 demonstrates that there is no significant difference of the RMSF profiles between the solution and the intercalation simulations with the new force field:

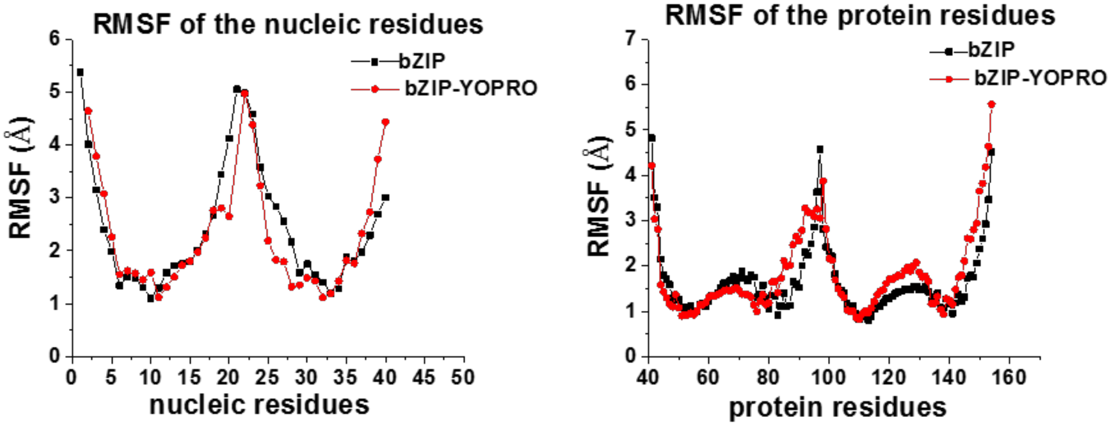

Figure 3.3.8 Model 2 RMSFs of the DNA bases and bZIP residues compared for bZIP-DNA (black) and bZIP-DNA with intercalated YOPRO (red) over their MD trajectories.

This 10 ns MD trajectory was started from −150 degrees for D1 and 0 degrees for D2. Figure 3.3.9 shows that D2 still fluctuates around 0 degrees, as in solution, while D1 is fixed at around −165 degrees. The behavior is significantly different from the solution model 2 YOPRO result shown in Figure 3.3.7. It indicates that intercalated YOPRO in this model is reasonably close to planar. Certainly much more so than in solution.

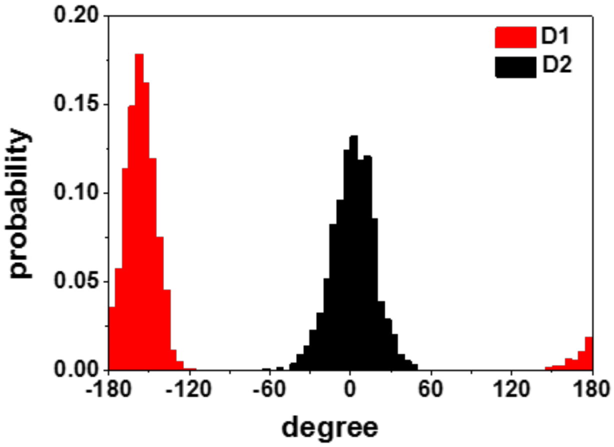

Figure 3.3.9 The dihedral angles D1 and D2 for intercalated YOPRO Model 2.

To confirm the conclusions regarding the stability of the YOPRO intercalation and its negligible effect on the structure of the bZIP-DNA-YOPRO complex the simulation was continued for another 13 ns. The results are presented as follows and are, to statistical fluctuations, the same as what has been presented in this section.

H bonds	Average bond lengths (Å)	Standard deviations (Å)
A37N1-T5N3	2.9473	0.1043
A37N6-T5O4	2.9983	0.1882
G38N1-C4N3	2.9530	0.08568
G38N2-C4O2	2.8702	0.1222
G38O6-C4N4	2.9232	0.1383

Table 3.3.7 The hydrogen bond average lengths and standard deviations over the 13 ns simulation. All H bond are well maintained.

Atom pairs	Average distances (Å)	Standard deviations (Å)
T5C5-C4C2	7.0020	0.7810
G38C4-T5C5	9.9039	0.8569
A37C2-C4C2	8.5856	0.4828
A37C2-G38C4	7.4347	0.4508
A37C2-T5C5	5.9951	0.1620
G38C4-C4C2	6.2891	0.1088

Table 3.3.8 The box atom pair average lengths and standard deviations over the 13 ns simulation.

Atom pairs	Average distances (Å)	Standard deviations (Å)
A37N7-YOPRO C3	4.0691	0.4155
C4C2-YOPRO C14	4.2586	0.5404
G38N3-YOPRO C8	5.2105	1.0278
T5C4-YOPRO C17	4.3699	0.5268

Table 3.3.9 The atom pair average lengths and standard deviations for the distances used to characterize the intercalation extent over the 13 ns simulation.

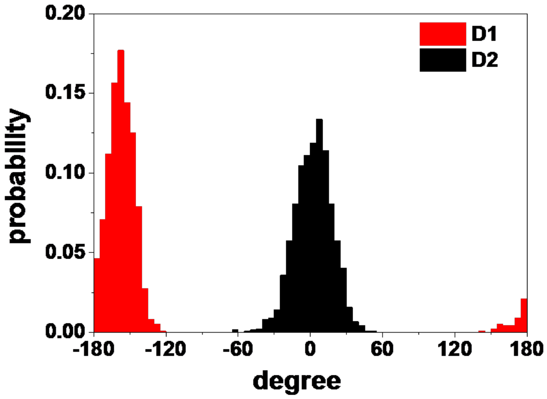

Figure 3.3.10 Model 2 extended dihedral angles D1 and D2 for intercalated YOPRO.

As for YOPRO model 1 in water, we carried out umbrella sampling to investigate this result. Of course, with an intercalated YOPRO, if a PMF around one or both linker dihedrals is constructed there is the distinct possibility that, for some twist angle range between the two ring systems, YOPRO will no longer be intercalated and/or the bases will no longer be hydrogen bonded, or form a box. The windows procedure was started from D1 at −152.4 and D2 at −4.6 degrees, where the probability is maximum. The D1 dihedral was restrained to cover the range (−180,180) while the D2 dihedral was restrained around 0 degrees (see Section 2.1). This probability is shown in Figure 3.3.11.

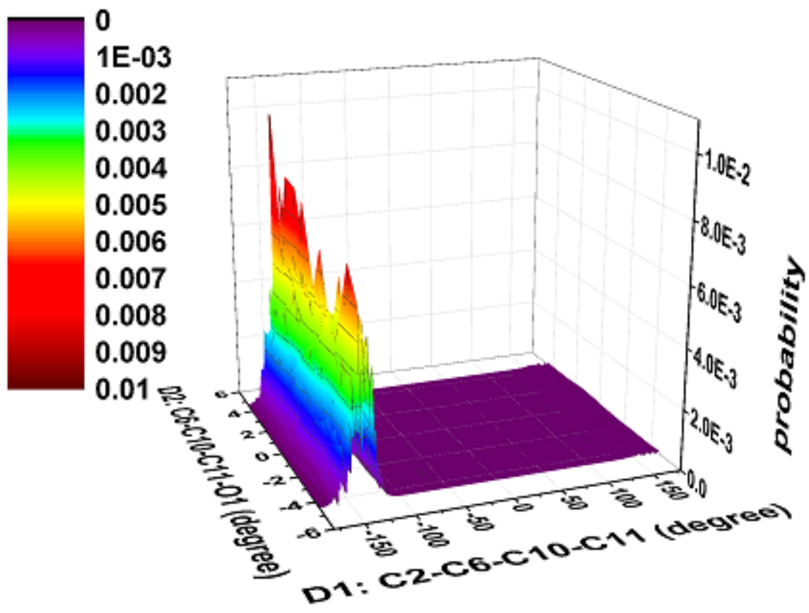

Figure 3.3.11 The intercalated YOPRO probability density (top) for the two dihedrals from a 2D umbrella sampling simulation.

The minimum on the PMF profile shifted from about –160 to –135 degrees. This is going away from the more planar conformation of the intercalated MD simulation. To understand this result, we explored the possibility that the ring system might no longer be well confined by the box formed from the surrounding bases. If, for example, the benzoxazole ring protruded from the box, its rotational motion could become unfrozen. As indicated in Figure 3.3.12, the G38N3-YOPRO C8 and A37N7-YOPRO C3 distances can be used to characterize the quinoline ring position relative to the bases, while the T5C4-YOPRO C17 and C4C2-YOPRO C14 distances can be used to characterize the benzoxazole ring position relative to the bases.

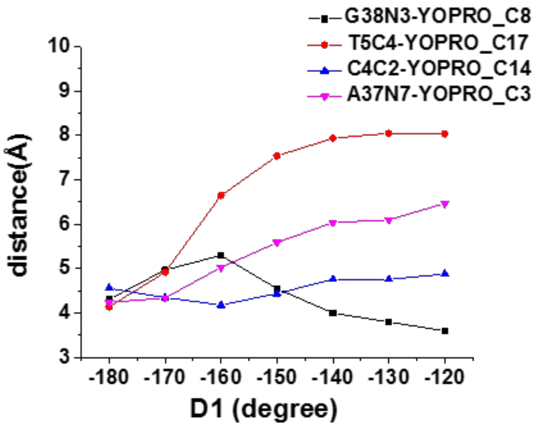

Figure 3.3.12 Four characteristic distances (see Figure 3.3.2) to monitor the positions of the ring systems, as a function of the D1 dihedral angle.

Figure 3.3.12 plots these four characteristic distances as a function of the D1 dihedral angle. The T5C4-YOPRO C17 distance rapidly increases when D1 goes beyond −160 degrees, indicating that the benzoxazole ring does go out of the box when the conformation of YOPRO is constrained to be further from planar.

The relative location of YOPRO shown in Figure 3.3.13 supports this conclusion.

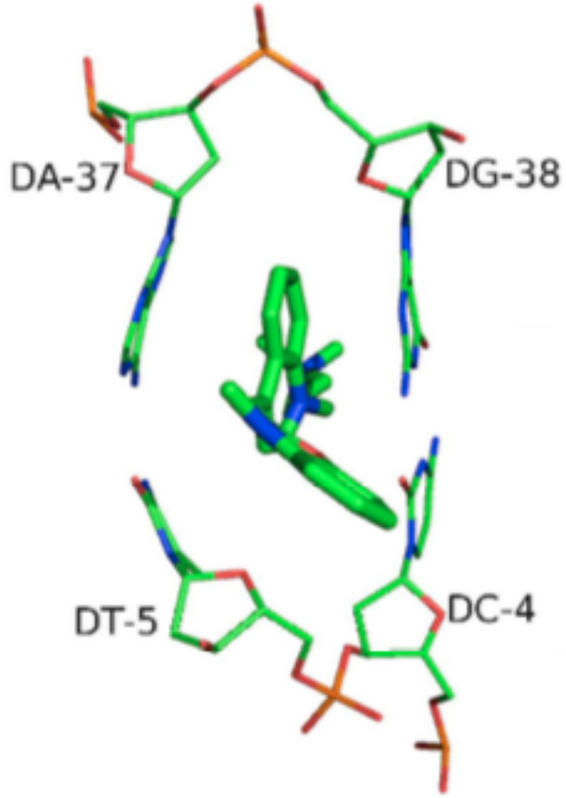

Figure 3.3.13 The last snapshot from the window where D1 is constrained at −120 degrees. In this twisted configuration the benzoxazole ring has moved out of the box formed from the surrounding bases.

It is significant that the five hydrogen bonds of the box base pairs are still well-preserved during the windowing, as shown in Figure 3.3.14. Thus, the box conformation didn’t undergo very significant changes when the benzoxazole ring departs from the box confines suggesting the existence of another binding mode.

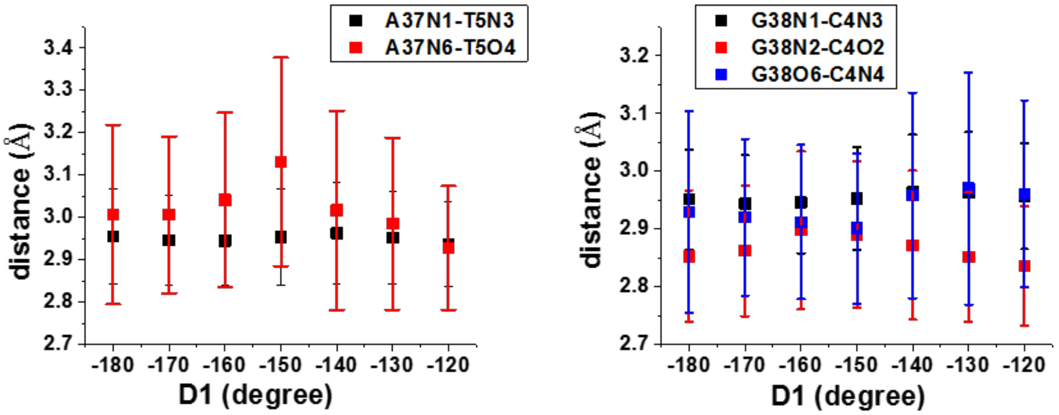

Figure 3.3.14 Model 2 average and standard deviations of the three AT and two GC hydrogen bonds as a function of the angle D1. The hydrogen bonding of the four bases is maintained even as the oxazole ring is no longer intercalated.

In addition, the RMSFs of the residues for the most twisted window, (–120, 0), also show that they are hardly affected by the benzoxazole de-intercalation as below.

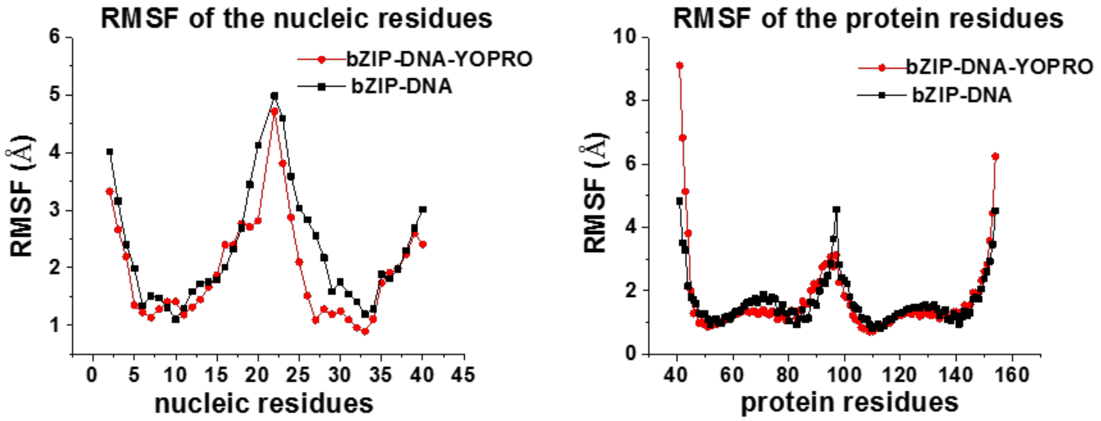

Figure 3.3.15 Model 2 extended RMSFs of the DNA bases and bZIP residues compared for bZIP-DNA and bZIP-DNA with intercalated YOPRO of the most twisted window, (-120, 0) over their MD trajectories.

3.3.7 YOPRO Model 3

That YOPRO is a conjugated system is quite clear from the quantum chemistry and from previous work on cyanine dyes as a class.[68-70] It is worth exploring a somewhat reduced value for these force constants to see what sensitivity there is to the bridging dihedral force constants. We use the values 0.5 kcal/mol for this model 3 as a value that is about the thermal energy.

3.3.8 YOPRO Model 3 in Water

YOPRO was simulated in water for 30 ns with the new force constant values. Figure 3.3.8.1 displays the two dihedral probability distributions.

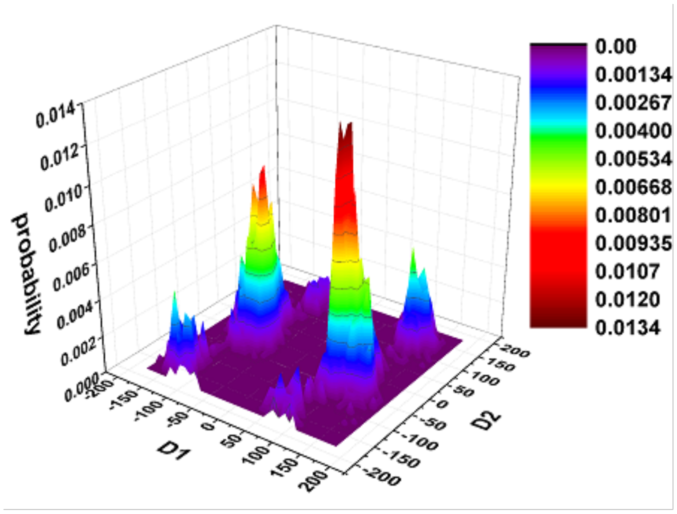

Figure 3.3.16 The dihedral angles D1 and D2 for unrestrained YOPRO Model 3 in water.

The D1 probability distribution is similar to Model 2 (see Figure 3.3.6) but without the "holes" in the distribution. D2 is still peaked around zero degrees. Thus there is some influence from the reduction in force constant, but the features that YOPRO in water is far from planar and samples broad ranges of these dihedral angles are still in evidence.

3.3.9 YOPRO Model 3 Intercalated

The bZIP-DNA complex with YOPRO intercalated was simulated for 13 ns. Monitors of the stability of Model 3 intercalation are presented in the figure 3.3.17 and the tables 3.3.10–3.3.12. Figure 3.3.18 displays the probability distributions of the dihedrals.

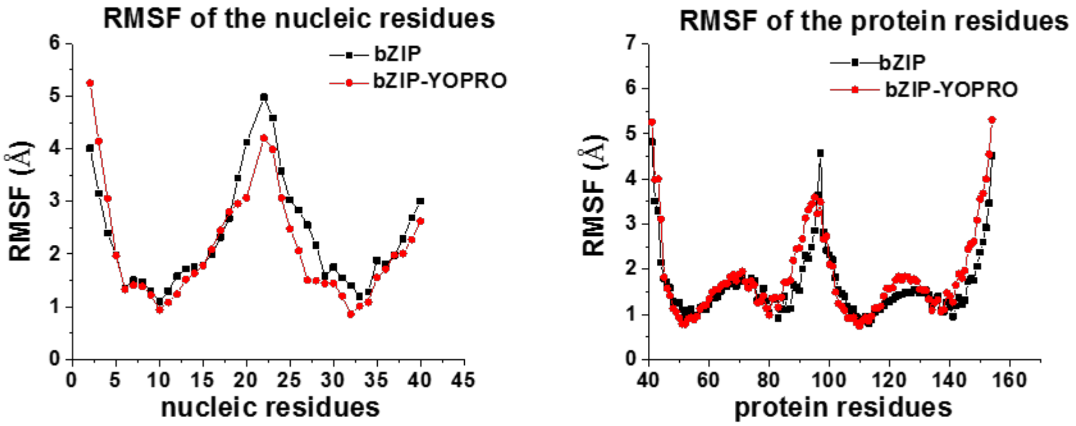

Figure 3.3.17 Model 3 RMSF of the DNA bases and bZIP residues compared for bZIP-DNA and bZIP-DNA with intercalated YOPRO over their MD trajectories.

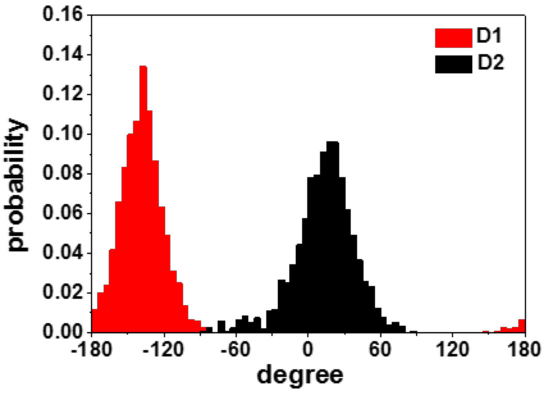

Figure 3.3.18 The dihedral angles D1 and D2 for intercalated YOPRO Model 3.

D1 is now centered at –140 degrees, while the equilibrium position of D2 is shifted to 20 degrees. Therefore the angle between the two ring systems is

$-140+20=-120$ degrees. That is a significant shift away from planarity in the Model 2 intercalation simulation. Furthermore, in contrast to Model 2, in this simulation, the oxazole ring does not remain intercalated (see Table 3.3.12). After about 3 ns over an interval of about 200 ps its position shifts from being intercalated to de-intercalated, much like in the Model 2 PMF simulation.

H bonds:	Average bond lengths (Å)	Standard deviations (Å)
G38O6-C4N4h:	2.918	0.1463
G38N2-C4O2h:	2.901	0.1338
G38N1-C4N3h:	2.953	0.09067
A37N6-T5O4h:	3.184	0.3313
A37N1-T5N3h:	2.962	0.1527

Table 3.3.10 The hydrogen bond average lengths and standard deviations over the simulation. All H bond are well maintained.

Atom pairs	Average distances (Å)	Standard deviations (Å)
A37C2-C4C2	8.3688	0.4334
A37C2-G38C4	7.6844	0.3977
A37C2-T5C5	5.9107	0.1753
G38C4-C4C2	6.2544	0.1143
G38C4-T5C5	9.4147	0.6400

Table 3.3.11 The box atom pair average lengths and standard deviations over the simulation.

Atom pairs	Average distances (Å)	Standard deviations (Å)
A37N7-YOPRO C3	5.5613	0.6491
C4C2-YOPRO C14	4.5867	0.5057
G38N3-YOPRO C8	4.1045	0.4293
T5C4-YOPRO C17	7.7156	0.4282

Table 3.3.12 The atom pair average lengths and standard deviations for the distances used to characterize the intercalation extent.

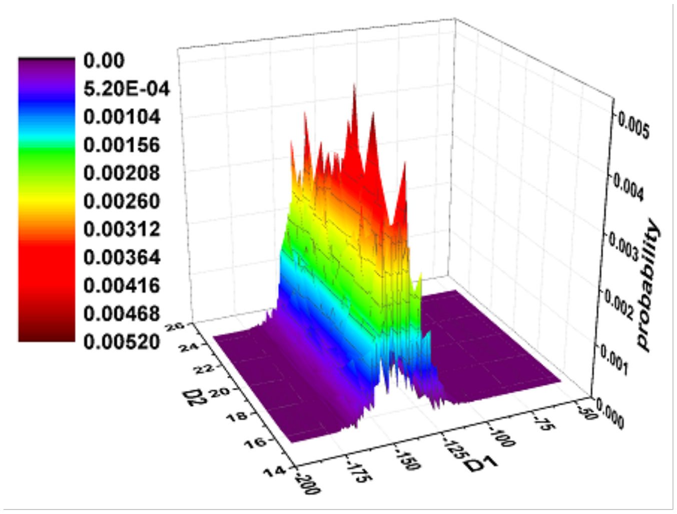

Figure 3.3.19 The intercalated YOPRO probability density for the two dihedrals from the 2D umbrella sampling simulation for YOPRO Model 3.

Similar to model 2, we also carried out umbrella sampling to explore the free energy surface around the probability maxima of the unrestrained simulation, with D1 around –140 and D2 around 20 degrees as shown in Figure 3.3.19. Windows were spaced 10 degrees apart for D1 while D2 was always constrained around 20 degrees. The details are given in the Section 2.1.

The free energy minimum is found around D1 = –135 degrees, which is consistent with the result from the unrestrained simulation. This is quite far from the planar conformation, indicating that the reduced force constants leads to deviations from planarity, compared to the original force constants 1 kcal/mol of Model 2. This PMF run was initiated from the end of the unrestrained run with the oxazole ring de-intercalated, and measurement of the characteristic distances shows that it remains de-intercalated.

3.3.10 Discussion and Conclusion Remarks of the Three Models

Three models of YOPRO were investigated. In model 1 the two linker bonds are treated as a single (C10-C11) and a double bond (C10-C11) with standard GAFF values for the dihedrals associated with these bonds. For models 2 and 3 both bonds are treated as appropriate to a delocalized bonds resulting in dihedral force constants that are intermediate between single and double bond values, as detailed in the Methods Section. Model 3 was generated by simply reducing the model 2 value by a factor of two. Thus, model 2 should be considered as the most accurate model based on the GAFF methodology.

The results of our simulations are summarized in Table 3.3.13. Only the peak positions are indicated, but looking at the various plots indicated in the table shows that the intercalated simulations lead to somewhat more constrained dihedral angle sampling than do the water simulations.

	Model 1 [a]		Model 2		Model 3	
	D1	D2	D1	D2	D1	D2
Water, unrestrained	−100 [b]		±70; ±140 [c]	0 [c]	±100 [d]	0 [d]
Water, window	−115[e]		±65; ±130 [f]	±10 [f]	–	–
Inter, unrestrained	−145 [g]		−165 [h]	0 [h]	−140 [i]	20 [i]
Inter, window	–		−135 [j]	–	−135 [k]	–

Table 3.3.13 Dihedral probability peaks (degrees) for all the simulations.
[a] Only D1 fluctuates. [b] Figure 3.3.2.1. [c] Figure 3.3.5.1. [d] Figure 3.3.8.1. [e] Figure 3.3.2.1. [f] Figure 3.3.5.2.[g] Figure 3.3.3.1. [h] Figure 3.3.6.2. [i] Figure 3.3.9.2. [j] Figure 3.3.6.4.[k] Figure 3.3.9.3

Model 1 in water shows that YOPRO is far from planar in the unrestrained simulation and is in good agreement with the umbrella sampling simulation over the range sampled by the unrestrained simulation. That the unrestrained sampling of D1 is so limited does show that even in water there are substantial barriers to this dihedral. Turning this dihedral introduces some van der Waals clashes within YOPRO, along with various solvation effects. When intercalated, D1's sampling shifts from peaking around–120 to around −145 degrees. Though the YOPRO intercalated between the base pairs is somewhat more planar than the unrestrained YOPRO in water, both are far from planar.

In model 2, the water unrestrained and window simulations qualitatively agree with each other. The water simulation has four peaks in the probability distributions. The D1 (C2-C6-C10-C11 dihedral) is always far from planar and the D2 (C6-C10-C11-O1) dihedral samples around zero degrees. The window simulation does put more of the D1 weight on the ±65 versus ±130 degree peaks than does the unrestrained simulation. In either case a large range of angles between the planes of the two rings is sampled.

The model 2 intercalated unrestrained simulation shows that the angle between the two ring systems is much closer to planar with (D1, D2) around (−165,0) degrees. The bases surrounding YOPRO are quite constraining though there are still fluctuations of both D1 and D2 around their average values. But the fluctuations are certainly much smaller than in water.

In the corresponding model 2 intercalation window simulation, where D2 was restrained around zero degrees and D1 rotated through its entire range, the possibility of de-intercalation of either one of or both ring systems, break-up of the four intercalating bases and general decomposition of the bZIP-DNA complex arises. What did happen, actually, is a shift of the D1 peak from −165 to −135 with a corresponding change in the position of the benzoxazole ring so as to protrude from its confining box (see Figure 3.3.13), as D1 was rotated over this range. The probability displayed in Figure 3.3.11 strongly favors this conformation. Note that D1 and D2 are, respectively, proximally associated with the quinoline and benzoxazole ring systems. Nevertheless it is the benzoxazole that comes out of its confining box. That is potentially due to the pyrimidine bases that are associated with the benzoxazole ring system (see Figure 3.3.13) presenting less ring surface area than do the purine bases. In spite of the benzoxazole de-intercalation, the hydrogen bonding of the four confining bases was not significantly altered (see Figure 3.3.14) for all the sampled D1 values indicating that this binding mode does not disturb the DNA pairing. The experimental linear dichroism data on YOPRO that suggested a surface binding mode at higher concentration was based on an excitonic[71] interaction mechanism that indicates sufficiently close multiple dye binding to the DNA. Clearly, in the MD with one YOPRO molecule, such effects are beyond the scope of this investigation. While, it should be understood that any MD force field used may not completely reflect reality, it is still of interest that an external binding mode of lower free energy can be found, at least in a reaction coordinate-based, window simulation.

Model 3 also has both linker bonds equivalent but with reduced (by a factor of two) dihedral force constant values relative to model 2. The values are not fundamentally based, but were introduced to explore the consequences of increasing the ability to twist around these dihedrals. In water, the D2 probability mainly peaks around zero degrees, and is essentially the same as model 2. Along the D1 dihedral coordinate, the four peak behavior of model 2 is smeared out into two peaks. Thus reduction of the dihedral potential, though intrinsically weak (from a barrier of 2 to 1 kcal/mol, so essentially thermal) has a modest but noticeable effect on the probability along the D1 coordinate. The sampling is very diffuse and certainly far from producing planar configurations.

When intercalated, the unrestrained model 3 simulation produces an average D1+D2 angle (essentially the angle between the planes of the two ring systems) of –120 degrees. This is substantially different than the much more planar arrangement in Model 2. This difference can be traced to the "spontaneous" de-intercalation that occurs rapidly in the unrestrained simulation. Once the benzoxazole ring is no longer confined by the surrounding bases, it becomes freer in its ability to rotate. An exploration of the PMF with D2 restrained around 20 degrees, its average position in the unrestrained simulation, and D1 rotating through its entire range finds the free energy minimum around –135 degrees. Thus, again a stable de-intercalated conformation is found.

In summary, model 2 that is best among the three from the point of view of force field development, leads to a view of the intercalation that is in good agreement with experimental data.[64, 71] Namely, YOPRO does intercalate in a cage of four bases with a relatively planar conformation of the two ring systems. Furthermore, the orientation of the YOPRO ring planes is essentially perpendicular to direction of the dsDNA. Another potential mode of YOPRO binding is found where the oxazole ring is de-intercalated. A surface binding mode is also inferred in the experimental data[71] at higher YOPRO concentrations.

3.4 Defects of the Three Models

By examining three distinct parameterizations of the methine linker differing in the dihedral force constants, we revealed considerable sensitivity of the dye's conformational behavior and intercalation stability to the parameterization of the dihedral potentials involved. This finding points to the improper description of YOPRO and other cyanine dyes by automated force field parameter-assigning tools and would be relevant for future MD studies of YOPRO and related dyes.

However, these modified Amber-Generated models have several severe defects. Firstly, these models are not parameterized from quantum chemical calculations but based on the empirical estimations on the delocalized system on the linker. Therefore, the choice of Model 2 as the most accurate parameterization may not be reasonable.

Most importantly, none of these models investigated the possible correlations between D1 and D2 in this conjugated system. As we show in Section 3.3, this turns out to be crucial to the correct description of the linker dihedral force field and has a major influence on the sampling in both water and when intercalated.

Secondly, only one planar conformation (trans-cis as shown in Figure 3.3.2) was used to initialize the simulations of the intercalation, which leads to the limitation of the sampling range of the intercalated YORPO. Intuitively, all four planar conformations should be considered equivalently as the initial state since every time YOPRO is "trapped" in one conformation by the box constraints and can hardly jump to another without the de-intercalation. So the chosen trans-cis conformation may not represent all intercalation cases.

Thirdly, the bZIP bound DNA may not be necessary for the intercalation simulations. When intercalated, the YOPRO doesn't interact with the bZIP directly. The DNA concentration assays using YOPRO doesn't require protein-bound DNA samples either[72].

Finally, all simulation were performed on the electronic ground state, while the relevant dynamics during fluorescence take place in the excited state. The conformational probability distribution for solvated and intercalated YOPRO must be evaluated. Therefore, later on we carried out charge and force field parameter fitting using the IPolQ method[11] as in Section 2.2 and reran the MD simulations on both S0 and S1 states with the corresponding fitted force field respectively.

3.5 S0 and S1 YOPRO Free Energy Surfaces in Water with Fitted Force Constant Parameters

Due to defects of the AMBER modified force field models as discussed in section 3.4, we developed our own force field models for both S0 and S1 electronic states of YOPRO starting from IPolQ charge fitting model[11]. The whole fitting procedure is detailed in Section 2.2. The following is the results from performing the simulations with the fitted force field models and closely follows the our paper *Molecular Dynamics of Oxazole Yellow Dye in its Ground and First Excited Electronic States in Solution and when Intercalated in dsDNA* published in 2017[73].

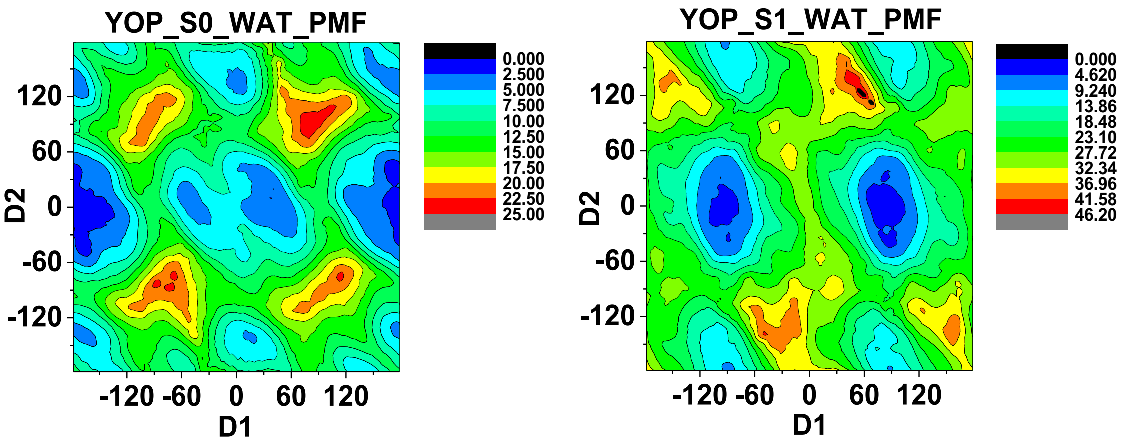

Figure 3.5.1 PMF surfaces of S0 and S1 YOPRO in water.

We carried out two-dimensional umbrella sampling to obtain free energy surfaces, as detailed in Section 2.2. The two dimensional free energy surfaces, also known as the potential of mean force (PMF) surfaces are defined as $PMF(D1,D2) \equiv -k_B T \ln[p(D1,D2)]$ where $p(D1,D2)$ is the corresponding probability distribution function. The results are shown in Figure 3.5.1:

The PMF plots of the S0 and S1 states reproduce the minima and maxima on the corresponding subtracted energy surfaces from subtracting the AMBER derived classical potential V from the quantum chemical energy Q as used for the force constant parameter fitting in Section 1.4 (Figure 3.5.2), which indicates that the dihedral potentials are quite strong compared to the solvation ability of water. On both the PMF and subtracted energy surfaces, S0 has four minima around (0, 0), (0, ±180), (±180, 0), (±180, ±180), which characterize the four planar conformations: cis–cis, cis–trans, trans–cis, trans–trans, respectively. In the S1 state, all of these minima become high-free-energy regions.

The probability analysis corresponding to the PMF shows that the S0 state is dominated by the trans–cis conformer, peaked around (±180,0). Thus, it should the region with the highest probability of Franck–Condon transitions. For the S1 state, the probabilities peak around (+90,0) and (−90,0).

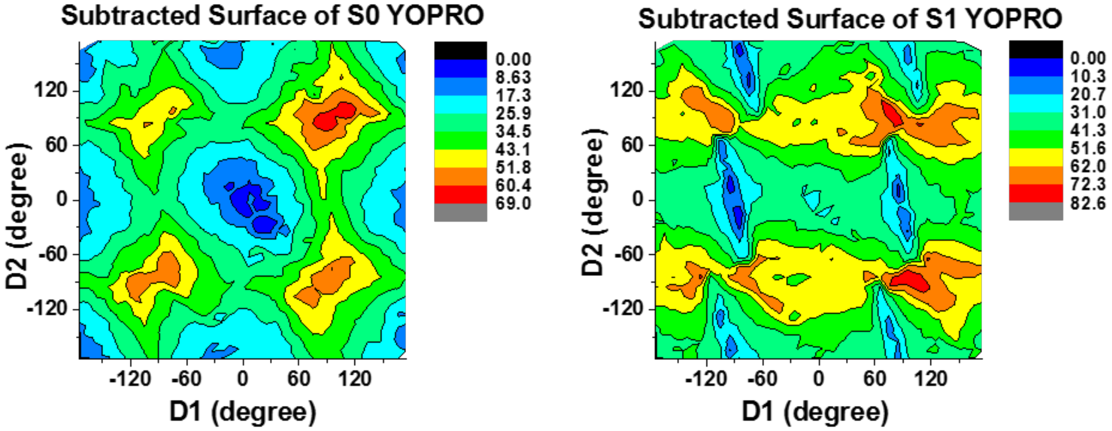

Figure 3.5.2 Subtracted energy surfaces of S0 and S1 YOPRO.

The umbrella sampling results of the S1 state in Figure 3.5.3, obtained by using the second method described in the Section 2.2, show that all four minima on the S0 surface correspond to high free energy regions of the S1 surface, which is consistent with the result of the full range sampling using the first method. Thus, they are also suitable starting points for obtaining trajectories on the S1 surface starting from the respective Franck–Condon points. However, as noted above, the S0 surface probability is dominated by the trans-cis conformation.

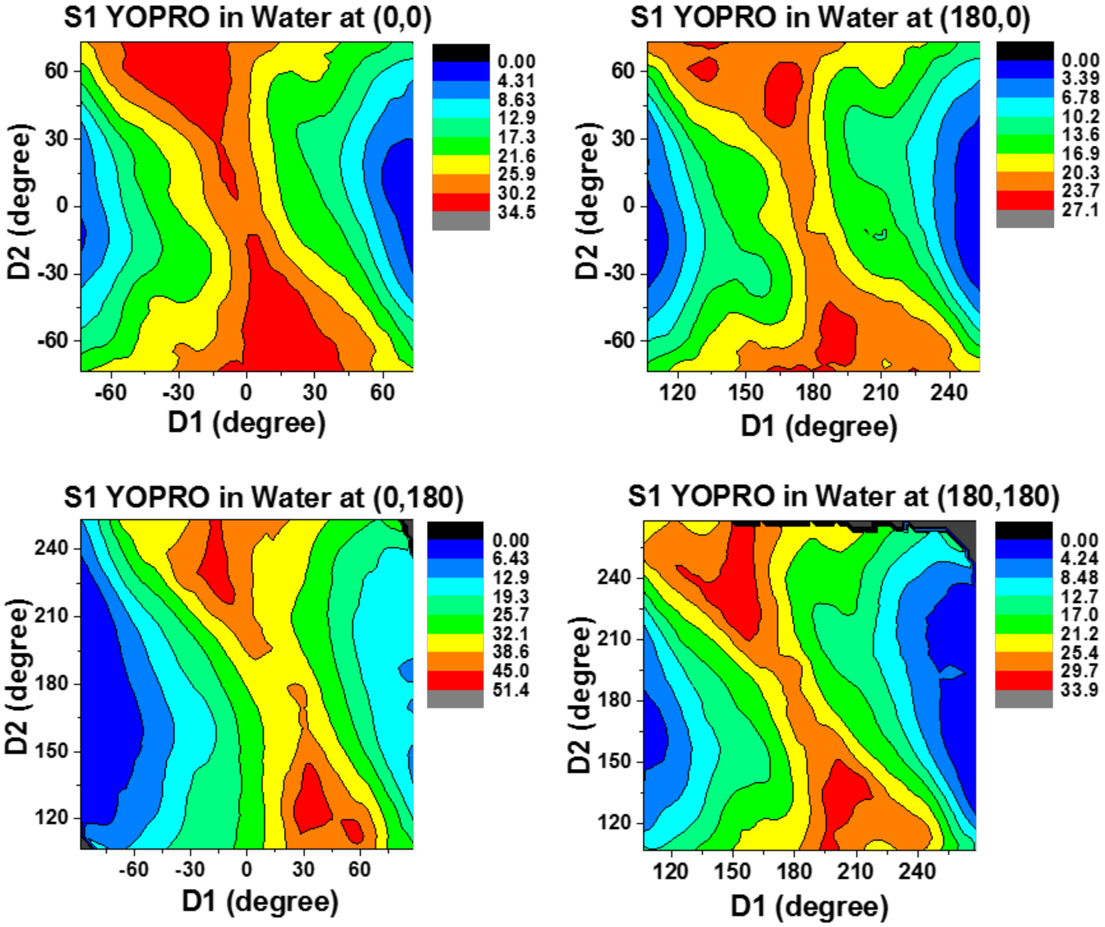

Figure 3.5.3 PMFs of S1 YOPRO in water around the four minima on the S0 quantum energy surface.

3.6 S0 and S1 YOPRO dsDNA intercalated.

As discussed in Section 2.2, we sampled intercalated YOPRO in the vicinity of the four local minima on the S0 state. The YOPRO does stably intercalate in the S0 state at all four planar conformations as in Figure 3.6.1. For the S1 state, PMFs of intercalated YOPRO around the three S0 minima at (0, 0), (0,180) and (180,180) show that intercalation of these planar conformations of S1 YOPRO doesn't change free energy profile pattern significantly.

The global minimum of the S0 state is at around (±180, 0) which, therefore, is the geometric origin of the Franck–Condon transition to the S1 surface. This most highly populated ground-state conformer de-intercalates when D1 goes below 150°. Thus, the only relaxation path for this S1-intercalated YOPRO is in the direction of increasing D1. The PMF surfaces of the two electronic states of the intercalations are shown below in Figure 3.6.1 and Figure 3.6.2.

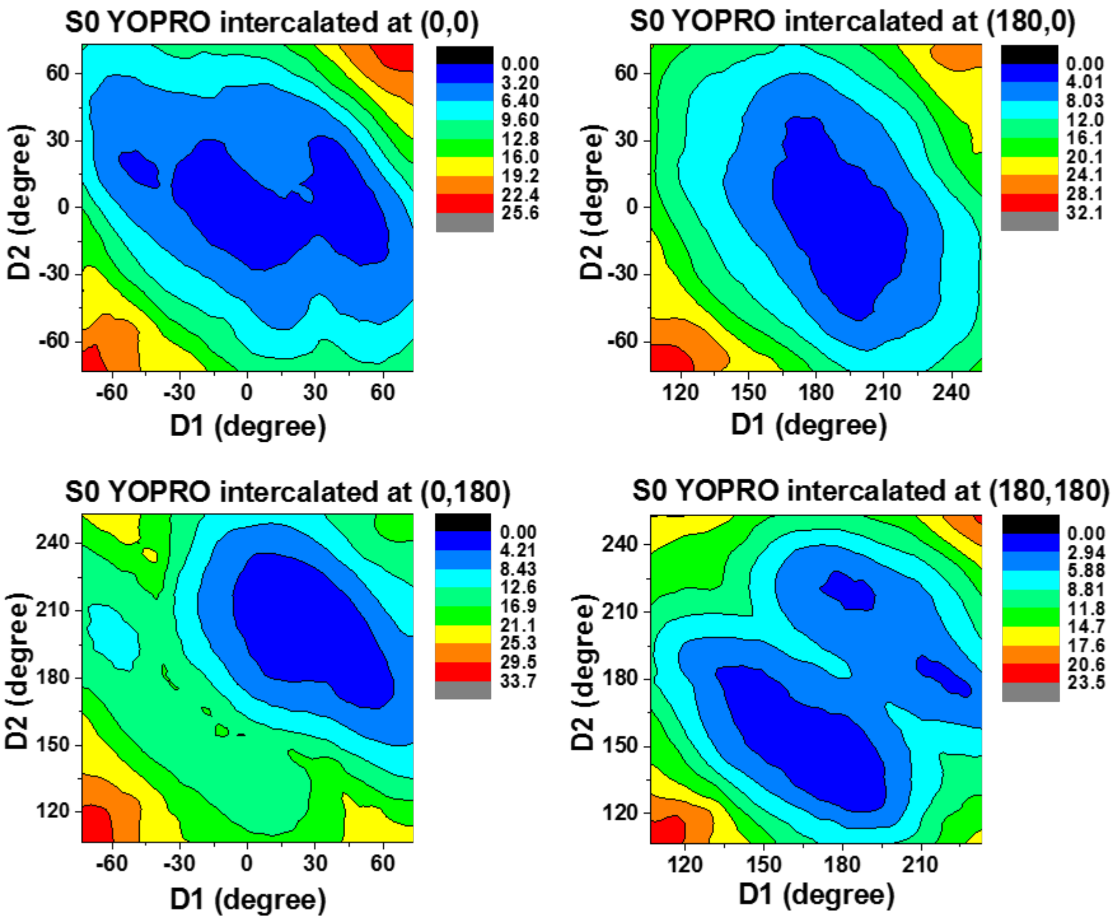

Figure 3.6.1 PMFs of intercalated S0 YOPRO centered at the S0 global minima.

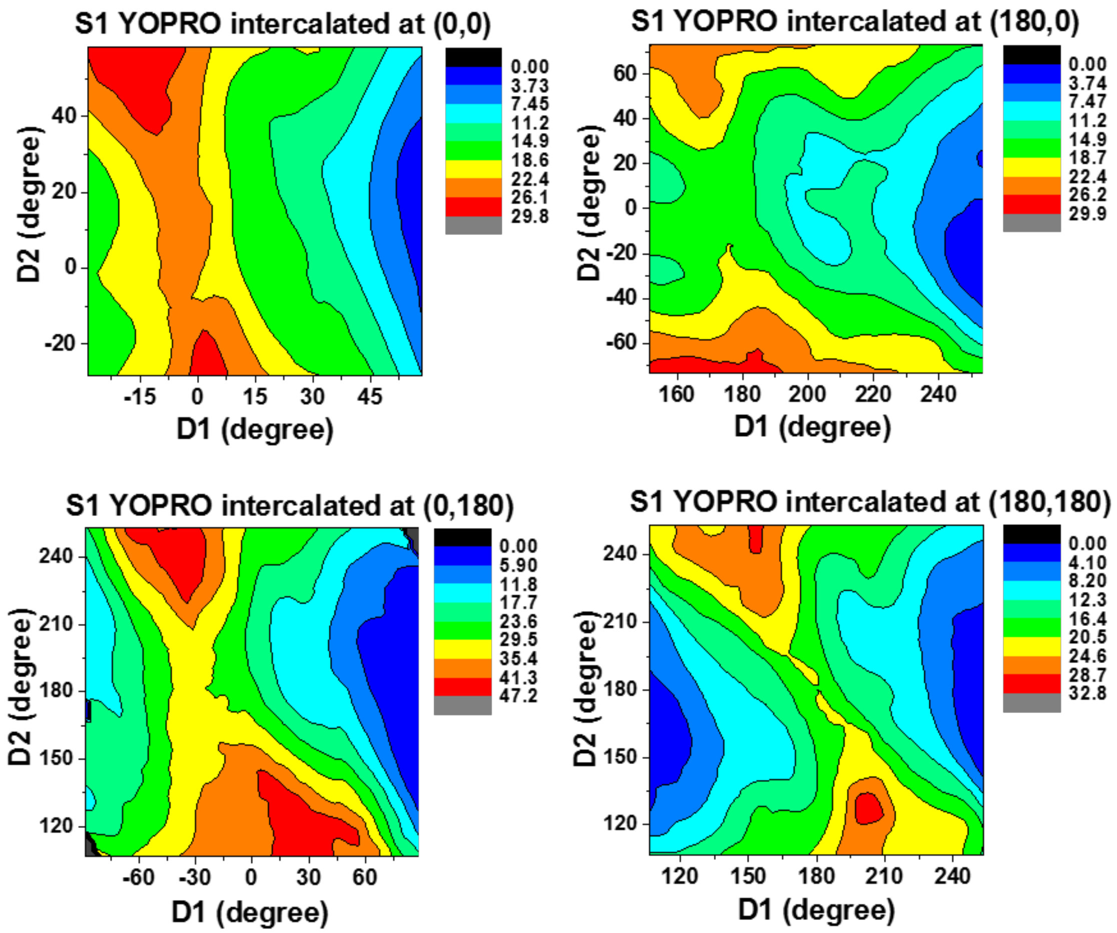

Figure 3.6.2 PMFs of intercalated S1 YOPRO centered at the S0 global minima.

3.7 Nonradiative Relaxation Path on the S1 Surface from the S0 Equilibrium Position

We obtained the relaxation pathway on the S1 surface of both water-solvated and intercalated YOPRO starting from the S0 equilibrium position (±180,0) using a steepest-descent algorithm that we implemented. As shown in Figure 3.7.1, the intercalated path has a distinct barrier between the initial (FC point) and final (local minimum) positions that is absent in the water simulation path.

Figure 3.7.1a shows that starting from the lower point of 15.85 kcal/mol on the intercalated surface the relaxation path goes uphill at D1 = 210° and forms a barrier of 2.73 kcal/mol, whereas the path on the water-solvated surface starts from the higher point of 20.07 kcal/mol and is completely downhill. Figure 3.7.1b shows that the relaxation path for the intercalation is elongated due to the uphill region, which indicates that a longer free-energy pathway is involved in the relaxation process on the intercalated surface than on the water-solvated surface.

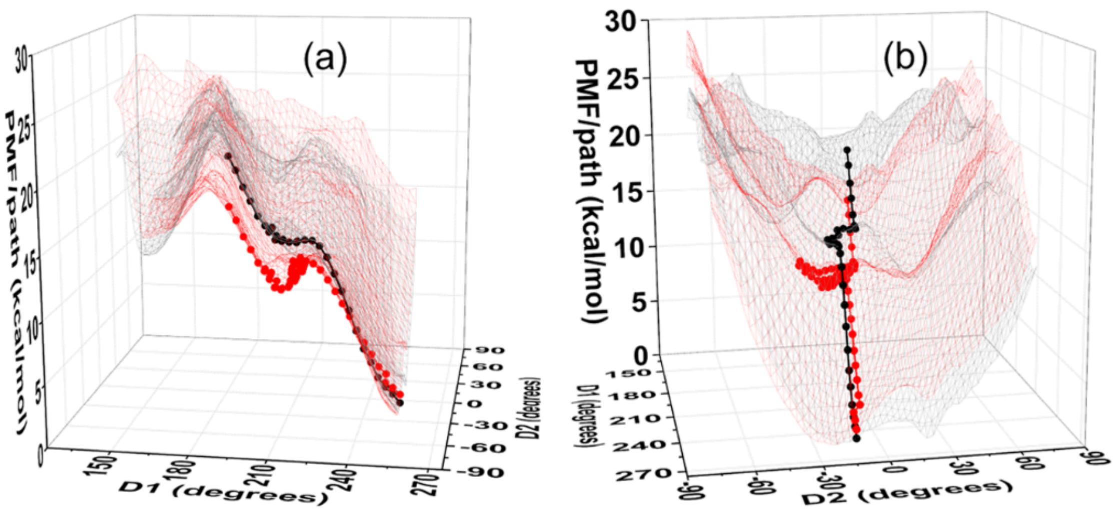

Figure 3.7.1 PMFs and relaxation pathways for S1 YOPRO in water and intercalated. The two views displayed, panels (a) and (b), are for better visualization. Black (path) and gray (PMF) are in water. Red (path) and light red (PMF) are intercalated. The barrier to relaxation is present only in the intercalated pathway.

3.8 Discussion and Concluding Remarks of the Results from the Fitted Force Field of the Two Electronic States

Conformational sampling of YOPRO in its S0 and S1 electronic states in solution and when intercalated in DNA was investigated with the aims of showing that YOPRO could stably intercalate in dsDNA and that intercalation leads to more constrained sampling around the linker dihedrals.

The Methods Section 2 used a recently developed procedure[11] to obtain potentials for the linker dihedrals. A coupling of the two dihedrals in both electronic states was observed. Therefore, an individual force constant fitting is essential for each dihedral conformational range. The S0 quantum energy surface along the two dihedrals shows a centrosymmetric pattern, which indicates the two C-C bonds on the linker (C6-C10 and C10-C11) are of the same bond order. This symmetry is not present in the S1 surface, revealing destruction of the conjugation of the linker due to charge transfer to the central carbon (C10) from its adjacent carbons (C6 and C11) [48]. Tables 2.2.3-2.2.4 show that in the S1 state the D1 dihedral (C6-C10) becomes more like a single bond when D2 goes to the high energy regions around ±90°, while the D2 dihedral (C10-C11) becomes more like a pure double bond when D1 goes to the same angles, which correspond to the energy minima.

The ground state dihedral population distributions in solution and when intercalated are essential to finding out where, in the vertical, Frank-Condon transition, there will be excited state population. Computational[48, 74, 75] and experimental[76-81] investigations of the S1 excited state of cyanine dyes suggest that torsional motions and other deformations in the bonds linking the two ring systems, if not restrained, can take place readily. From the computational perspective, there is consensus that the S1 potential surface varies more slowly than the S0 surface. Experiments suggest correspondingly rapid motions along the S1 surface. Thus, without some constraint mechanism, twisting in the excited state is facile, avoided crossings or conical intersections (CIs) between excited and ground electronic states are readily reached, and the ground state is accessed via a nonradiative pathway. This process leads to a great reduction in fluorescence intensity. To prevent twisting, various constraints have been invoked. Lowering the temperature and increasing solution viscosity, or both, are the standard ways to prevent access of nonradiative pathways.[82 83]

Dye Intercalation in DNA is another method of enforcing a constraint. Our umbrella samplings show that when intercalated in the S0 equilibrium conformation (trans-cis), the difference between the starting pointing, which is the free energy maximum on the S1 surface in the sampling region, and the local minimum, is reduced from 20.07 kcal/mol to 15.85 kcal/mol. On the S1 surface of ME-1122P, a related dye, the local minimum is in the vicinity of the conical intersection point.[74] This reduction is also observed when the S1 YOPRO is intercalated in the other three planar conformations (Figures 3.5.3 and 3.6.1), which correspond to the other low free energy regions on the S0 surface.

The S1 surface dynamics of starting from the Franck-Condon point to a conical intersection point that is in the vicinity of S1 surface minimum can be thought of as consisting of two steps: a diffusive motion along a reaction coordinate, followed by the transition from the S1 to S0 surface around the CI point. Thus, the overall rate constant for this consecutive process can be written as $k^{-1} = k_D^{-1} + k_R^{-1}$ with k_D a diffusive rate constant and k_R the rate of S1→S0 deactivation. Assuming that k_R is large compared with k_D as is often the case[83], the overall rate is dominated by k_D. For the intercalation case, where there is a barrier to transition, and what amounts to a high viscosity for the reorientation of the linker dihedrals, a Kramers rate expression in the highly overdamped regime is indicated.[84] In this regime, k_D is given by the expression:

$$k_D = \frac{\omega_0 \omega_b}{\gamma} e^{-E_b/k_B T} \tag{3.8.1}$$

where $\omega_0 \omega_b$ is the product of the well ω_0 and barrier ω_b frequencies, respectively, γ is the friction coefficient along the reaction coordinate and E_b is the barrier height relative to the well origin. The Boltzmann factor rate reduction from the barrier of 2.73 kcal/mol, when D1 goes to 210° and D2 fluctuates around 0°, would reduce the rate constant of non-radiative relaxation by about 100 fold. It is not possible to obtain a value for the friction coefficient for twisting of intercalated YOPRO, however it is safe to assume that it would be considerably larger than that for YOPRO twisting in water. Furthermore, the longer path to climb the barrier, when intercalated, also should decrease the rate of passage. Therefore, intercalation provides a combination of increased viscosity and a barrier to reaching the CI point, relative to the completely downhill process in water, and leads to the increase of fluorescence intensity with YOPRO intercalation.

3.8.1 C++ program implementing the steepest descent algorithm

```
#include<iostream>

#include<string>

#include<vector>

#include<fstream>

#include<utility>

using namespace std;

int main()

{

   ifstream input;

   string in_name = "";

   cout<<"input a file"<<endl;

   cin>>in_name;

   input.open(in_name);

   if(!input.is_open())

   {

      cout<<"No such file!"<<endl;

      return 0;

   }


   double xbsize,ybsize=0;

   cout<<"input xbin size, ybin size:"<<endl;

   cin>>xbsize>>ybsize;
```

```
double xmax,xmin,ymax,ymin =0;
cout<<"input xmin,xmax,ymin,ymax"<<endl;
cin>>xmin>>xmax>>ymin>>ymax;

int xbnum = 1+(xmax-xmin)/xbsize;
int ybnum = 1+(ymax-ymin)/ybsize;

double** PMF = new double* [xbnum];
for(int i=0;i<xbnum;i++)
{
   PMF[i] = new double[ybnum];
}

string skip="";
getline(input,skip);

for(int i=0;i<xbnum;i++)
{
   for(int j=0;j<ybnum;j++)
   {
      input>>skip>>skip>>PMF[i][j]>>skip;
      //cout<<PMF[i][j]<<endl;
   }
```

```
}

double x_st,y_st=0;

cout<<"starting point:"<<endl;

cin>>x_st>>y_st;

int x_index = (x_st-xmin)/xbsize;

int y_index = (y_st-ymin)/ybsize;

double PMF_st = PMF[x_index][y_index];

ofstream output(in_name+"result.txt");

vector<pair<int,int>> x_y;

x_y.push_back(make_pair(x_index,y_index));

cout<<xmin+x_index*xbsize<<" "<<ymin+y_index*ybsize<<" "<<PMF_st<<endl;

output<<xmin+x_index*xbsize<<" "<<ymin+y_index*ybsize<<" "<<PMF_st<<endl;


double neighbor_up = PMF[x_index][y_index];

double neighbor_down = PMF[x_index][y_index];

double neighbor_right = PMF[x_index][y_index];

//double PMF_next = PMF[x_index][y_index];


while(x_index<=xbnum-1 && x_index>=0 && y_index<=ybnum-1 && y_index>=0)

{

  if(y_index<ybnum-1)

  neighbor_up = PMF[x_index][y_index+1];
```

```
    if(y_index>0)
    neighbor_down = PMF[x_index][y_index-1];

    if(x_index<xbnum-1)
    neighbor_right = PMF[x_index+1][y_index];
    else
    break;

    double PMF_old =PMF[x_index][y_index];
    double PMF_next = PMF[x_index][y_index];

    if(neighbor_up<=neighbor_down && neighbor_up<=neighbor_right /*&&
neighbor_up<=PMF_old*/)
    {
      y_index++;
      if(x_y.size()>2)
      {
        if(y_index==x_y[x_y.size()-2].second && x_index == x_y[x_y.size()-2].first)
//prevent stepping back and forth
        {
          y_index--;
          if(neighbor_right<=neighbor_down)
          {
              PMF_next = neighbor_right;
              x_index++;
          }
          else
          {
              PMF_next = neighbor_down;
              y_index--;
          }
          if(PMF_next==PMF_old)
          {
              cout<<"done"<<endl;
              break;
          }
          x_y.push_back(make_pair(x_index,y_index));
          cout<<xmin+x_index*xbsize<<" "<<ymin+y_index*ybsize<<"
"<<PMF_next<<endl;
          output<<xmin+x_index*xbsize<<" "<<ymin+y_index*ybsize<<"
"<<PMF_next<<endl;
          continue;
        }
      }
      PMF_next = neighbor_up;
    }
```

```
    if(neighbor_down<=neighbor_up && neighbor_down<=neighbor_right /*&&
neighbor_down<=PMF_old*/)
    {
      y_index--;
      if(x_y.size()>2)
      {
        if(y_index==x_y[x_y.size()-2].second && x_index == x_y[x_y.size()-2].first)
//prevent stepping back and forth
        {
          y_index++;
          if(neighbor_right<=neighbor_up)
          {
            PMF_next = neighbor_right;
            x_index++;
          }
          else
          {
            PMF_next = neighbor_up;
            y_index++;
          }
          if(PMF_next==PMF_old)
          {
            cout<<"done"<<endl;
            break;
          }
          x_y.push_back(make_pair(x_index,y_index));
          cout<<xmin+x_index*xbsize<<" "<<ymin+y_index*ybsize<<"
"<<PMF_next<<endl;
          output<<xmin+x_index*xbsize<<" "<<ymin+y_index*ybsize<<"
"<<PMF_next<<endl;
          continue;
        }
      }
      PMF_next = neighbor_down;
    }
    if(neighbor_right<=neighbor_up && neighbor_right<=neighbor_down /*&&
neighbor_right<=PMF_old*/)
    {
      x_index++;
      PMF_next = neighbor_right;
    }
    if(PMF_next==PMF_old)
    {
      cout<<"done"<<endl;
      break;
    }
```

```
        x_y.push_back(make_pair(x_index,y_index));
        cout<<xmin+x_index*xbsize<<" "<<ymin+y_index*ybsize<<"
"<<PMF_next<<endl;
        output<<xmin+x_index*xbsize<<" "<<ymin+y_index*ybsize<<"
"<<PMF_next<<endl;
    }
}
```

4 Protein Classification Using Deep Neural Networks

4.1 Introduction

Proteins are functional macromolecules for many biological processes. They consist of one or more peptide chains of amino acid residues that are folded in the three dimensional space. A peptide chain is usually synthesized by the dehydration of twenty types of amino acids and their derivatives, forming the primary structure of a protein. The chain is then folded into the secondary structures such as alpha helices and beta sheets by forming hydrogen bonds between non-adjacent amino acids. One or more secondary structures grouped together form the tertiary structure of a protein by covalent interactions between the side chains of the neighboring amino acids, for example, a disulfide bond formed between two cystines. The tertiary structure is the geometric structure of a protein. Finally, a number of tertiary structures as subunits can be folded into the quaternary structure of a multi-subunit protein complex like hemoglobin and ion channels.[85, 86]

Proteins are classified into different families. A protein family is a group of proteins that share a common evolutionary ancestor and usually a similar primary structure and physiological functions. There are many protein family databases. For instance, Pfam[12], collects 16,712 protein families in the most recent vesion. For each family, a representative set of sequences are aligned into a seed alignment, which is then used to build a profile hidden Markov model (HMM)[87, 88]. The HMM is then searched against the sequence databases and classify all hits reaching a manually curated gathering threshold into that family. The collection of the hits is then aligned to the HMM to generate a full alignment. Another example is UniProt[89], which was built on the collaboration of Swiss Institute of Bioinformatics, European Bioinformatics Institute and Protein Information Resource.[90-94]

There are other more machine learning oriented ways to predict whether a given protein belongs to a certain family or not by using the sequential information of the peptide chains directly or indirectly. For example, Asgari and Mofrad developed the ProtVec model[95] which embeds 3-grams of amino acids, namely the sequence consisting of any three neighbor residues, into a vector of length 100. Such a vector is obtained in the same way as the word representations[96] that are used in a skip gram model[97] by optimizing the softmax function of the 3-gram:

$$\frac{1}{N}\sum_{i=1}^{N}\sum_{-c\leq j\leq c, j\neq 0} p(w_{i+j} \mid w_i) \tag{4.1.1}$$

where w_i is the ith word in the text and 2c is the window size as the context that is centered at word i, and

$$p(w_{i+j} \mid w_i) = \frac{\exp(\mathbf{v}'^{T}_{w_{i+j}} \mathbf{v}_{w_i})}{\sum_{k=1}^{w} \exp(\mathbf{v}'^{T}_{w_k} \mathbf{v}_{w_i})} \tag{4.1.2}$$

where $\mathbf{v}_\mathbf{w}$ and $\mathbf{v}_\mathbf{w}'$ are the input and output vector representation of word w.

Each sequence is therefore represented by the element wise summation of its ProtVecs of all 3-grams. This method makes use of only the primary structure information but ignore the higher order structure information. In this chapter, we aim to predict protein families using both the primary structure and the tertiary structure information.

Our plan is to train three models: 1) a vanilla neural network with one hidden layer 2) a convolutional network 3) a long short-term memory network. The first two models use only the sequential information of the protein sequences and the resulting prediction accuracies are compared, and the last one uses both the sequential and the structural information as a comparison of the accuracies.

4.2 Data Sources

For the first two models, the input data comes from the fasta files downloaded from the UniProt website (https://www.uniprot.org/uniprot/)[89]. A fasta file is a collection of discovered sequences of a protein family in which the amino acids are represented using single-letter codes. Four families are picked for their abundance of sequences:

1. PDOC00137 (ATP synthase alpha and beta subunits)
2. PDOC50003 (GTPASE PH domain)
3. PS00178(Aminoacyl-transfer RNA synthetases class-I)
4. PS50862(Aminoacyl-transfer RNA synthetases class-II)

From each family we picked 7500 sequences with the total length (number of residues) falling between 600 and 800. The first 600 residues of a sequence are used as an input sample. The one letter codes of the residues are converted into their alphabetic indices as the numeric inputs. For instance, alanine has "A" as its one letter code, therefore it's represent by the integer "0"; Asparagine's one letter code is "N", which is converted into the integer "1". Among these 15000 sequences in total from the two classes we randomly picked 10500 for training and all rest as the testing set.

For the third model that also uses structural information, the input data are extracted from the protein data bank (PDB) files on the protein data bank website[98] (https://www.rcsb.org/). A PDB file contains, in addition to sequence, the coordinates of all atoms in the molecule. We picked another four classes each of which provides 400 PDB files:

5. Helicase

6. Transmembrane receptor

7. Amino acid kinase

8. ATPase

For each protein sample, the first 100 residues which is the approximate average length of a protein domain are truncated as an input. For the both sequential and structural inputs, each residue is represented by an integer vector of length 26 which is the Alphabet length.

For the sequential inputs, the representation of a residue is a one-hot vector that has one at the alphabetic index of its one letter code and zeros for all other elements. For example, the one letter code of cysteine is “C”, so it’s an one-hot vector started with two “0”s followed by one “1” and twenty five “0”s, that is: [0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0].

For the structural inputs, we tested two ways to represent a residue:

1) the vector of a residue has one at the alphabetic indices of its own one letter code as well as its neighbors, and zero for all the others. A neighbor is another residue inside its “neighborhood” which is the region of a certain radius centered at its Cα atom. For example, if an leucine (L) has two alanines (A) and one arginine(R) in its neighborhood, it’s vector representation would have a one for the 11th element (contributed by itself), a one for the first element (contributed by the two alanines), a one for the 17th element (contributed by the arginine) , and zeros for all the others, that is, [1,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0].

2) The vector records the numbers of the type occurrences of the residue’s neighbors. The vector representation of the example mentioned above now has a two

for its first element, that is, [2,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0]

4.3 List of terms used

This list provides the definition of frequently terms in machine learning:

artificial neural network: abbreviated as neural network, it's a collection of layers of nodes, known as artificial neurons, that is used to transmit the input features of the samples and output their label predictions.

feature vector: vectorized representation of a sample. For a residue, it is an integer of length 26; for a whole sequence, it's composed by concatenating the feature vectors of all its residues.

input matrix: a matrix consisting of feature vectors. The number of rows is the number of input samples and the number of columns is the length of the feature vector.

parameter matrix: a matrix to which a feature vector or an input matrix is multiplied to generate the predicted results.

parameter vector : if the predict results are scalars, then the parameter matrix is reduced into a parameter vector with only one column.

label: desired output of a sample. For example, the family that a protein belongs to.

loss function: the function that measures the extent to which the predictions deviate from the observed labels.

activation function: a function acting on the output of a layer element-wisely to add nonlinearity to the neural networks. The sigmoid function and the rectified linear unit (ReLU) function as mentioned below are two typical activations functions.

gradient: first order derivative of the loss function with respect to the parameters.

learning rate: the step size by which the parameter matrix elements are updated at each iteration.

bias: the term added to the product of the parameter matrix and feature vector to make the prediction closer to the label

overfitting: the trained model corresponds too closely to the training set and therefore fails to predict the testing data well.

regularization: a technique to relieve overfitting by minimizing the sum of the loss function and the norm of the parameter matrix together.

filter: a matrix that is used to convolve the input data.

stride: the step size by which a filter moves each time.

accuracy: number of true positives and true negatives divided by the total number of samples.

Some other specific terms are defined with equations in the following sections.

4.4 Neural Network Experiments

4.4.1 Model 1: Vanilla neural network with one hidden layer

For the first model, we implemented a vanilla network[13]. This network has one input layer which is the input matrix(10500 samples × 600 features), followed by a hidden layer of 40 neurons, and an output layer of one neuron indicating what family the protein belongs to. The loss function is defined as below:

$$J(\mathbf{w}) = y\log(\frac{1}{1+e^{-\mathbf{w}^{\mathrm{T}}\mathbf{x}}}) + (1-y)\log(\frac{1}{1+e^{-\mathbf{w}^{\mathrm{T}}\mathbf{x}}}) \tag{4.4.1}$$

where **w** is the parameter vector mapping the feature vector to the label, **x** is the feature vector and y is the label of a particular sample.

The learning rate is set to 0.003, meaning that for each iteration, the parameter matrix elements are updated by -0.003×gradient. The gradient vector is the first order derivative of the loss function with respect to the parameter vector. We used 30000 iterations for the whole training process.

This network is designed for binary classifications[99], namely to classify the samples into two families. Therefore, each time two classes are picked. The accuracies shown in Table 4.4.1 are defined as the sum of all true positive and true negative predictions divided by the total number of the test samples:

training/testing accuracies	class 1	class 2	class 3
class 2	0.999/0.903	-	-
class 3	0.994/0.926	0.988/0.860	-
class 4	0.996/0.927	0.987/0.862	0.992/0.888

Table 4.4.1 Accuracies of binary classification using the vanilla neural network

The difference between the training and testing accuracies indicates the existence of the overfitting issue[100], since for each pair, the resulting testing accuracy is lowered by about 10% compared to the training accuracy

The most commonly used technique for relieving overfitting is regularization.[101] It adds the L_2-norm of the parameter matrix w multiplied by a constant λ to the J(**w**) in equation 4.4.1, therefore,

$$J(\mathbf{w}) = y\log(\frac{1}{1+e^{-\mathbf{w}^{\mathbf{T}}\mathbf{x}}}) + (1-y)\log(\frac{1}{1+e^{-\mathbf{w}^{\mathbf{T}}\mathbf{x}}}) + \lambda\mathbf{w}^{\mathbf{T}}\mathbf{w} \qquad (4.4.2)$$

where λ is the coefficient used to determine the extent to which the norm of w contributes to the object J(**w**). When λ goes to 0, equation 4.4.2 degrades to equation 4.4.1.

We then tested regularization on class 2 and class 3 where the largest gap between the training and testing accuracy appears:

lambda	0.1	1	10	100	1000	10000
training/testing accuracies	0.985/ 0.859	0.988/ 0.861	0.987/ 0.860	0.832/ 0.757	0.921/ 0.843	0.718/ 0.711

Table 4.4.2 Accuracies of classifying class 2 and 3 with regularization.

According to the results above, the regularization didn't seem to make a big difference to relieve the overfitting.

Overfitting can be reduced by increasing the training set size and by decreasing what is included in the model. However, is seems better to change the NN method, as now discussed.

4.4.2 Model 2: Convolutional Neural Network

Our motivation for the second model is to improve the performance of the neural network by using a Convolutional Neural Network (CNN).[102]

In a CNN, a input sample is usually one-dimensional or two-dimensional, though it can be of higher dimension. Figure 4.4.1 shows a 4×5 matrix as a two-dimensional input (blue grid in the left panel). In the next convolutional layer, a convolution operation is applied to the input with a filter (red grid on the left panel). The part on the input covered by the filter is multiplied by the filter element wisely, and the products are summed to give a output square (red grid in the middle). The stride is set to 1×1, meaning that every time the filter goes to the right or to the bottom by a step of one element. Therefore, the output layer is a 3×4 matrix. Finally, a flatten operation concatenates all rows of the output matrix and makes a vector of length 12.

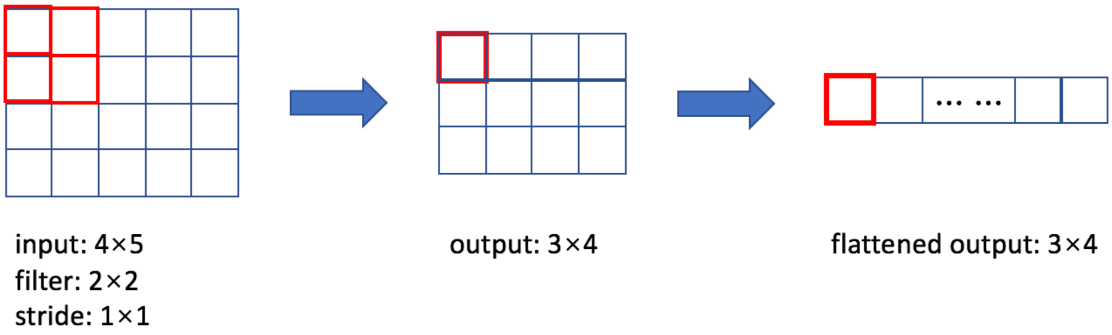

Figure 4.4.1 Convolution and Flatten Operation

Other convolutional operations are known as Batch Normalization and Max Pooling. In a Batch Normalization, what's covered by a filter is defined as a batch, and all values in the batch are subtracted by their average. While in a Max Pooling, the output of a convolution is set to be the maximum value of the batch convolved.

In this simple convolutional neural network model(CNN), we use the same input as for the vanilla network. The samples are reshaped into a 600×1 matrices to fit in the built-in Conv1D layer in keras[103], which is a framework written in Python for neural network designs. 64 filters of size 7 and stride 2 are used. After the convolution, a BatchNorm step is added to speed up the training.

Next, the rectified linear unit (ReLU)[104] activation function is applied, which takes the positive part of the Batch Norm result:

$$f(x) = \max(0, x) \tag{4.4.3}$$

It is followed by a MaxPooling layer using a filter of length 3 and stride of length 2. The inputs are then flattened and go to the output layer with a normalized exponential activation function as in equation 4.4.4:

$$s(x)_j = \frac{e^{x_j}}{\sum_{k=1}^{N} e^{x_k}} \tag{4.4.4}$$

The whole architecture of the model is described in Figure 4.4.2:

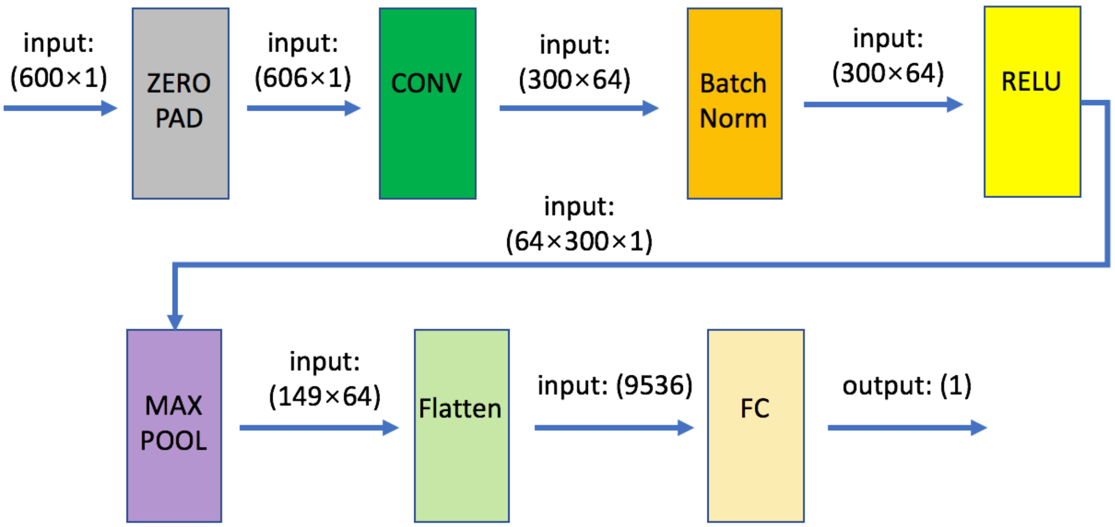

Figure 4.4.2 A simple CNN network. What's inside the parentheses above the arrows show the shapes of the sample inputted into the next step.

The accuracies are shown in Table 4.3:

training/testing accuracies	class 1	class 2	class 3
class 2	0.982/0.941	-	-
class 3	0.991/0.967	0.924/0.883	-
class 4	0.985/0.965	0.941/0.900	0.961/0.917

Table 4.4.3 Accuracies of using the above convolutional neural network

Apparently the convolutional neural network has a significantly improved performances. The testing accuracies are significantly increased although still lower than the corresponding training accuracies as expected. This results from the fact that convolutions reduce the dimensionality of the inputs, therefore relieves the overfitting issue.

4.4.3 A Deeper CNN

Although the simple 1D CNN with one hidden layer shows a much better performance than the vanilla neural network, we still want to test a deeper one. The main benefit of a deep network is that it can represent complicated nonlinear functions. It also learns features at many different levels of abstraction. Consider face recognition as an example: a deeper network learns the edges at the lower layers, and complex features like eyes and nose at the deeper layers.[105]

In a network of multiple layers, the inputs are propagated from the input layer through the hidden layers to the final output layer. This process is named forward propagation[106], each step of which is a matrix multiplication of the output matrix of the previous layer with the parameter matrix and the current layer:

$$\mathbf{z}^{(\mathbf{l+1})} = \Theta^{(\mathbf{l+1})} a^{(l+1)}(\mathbf{z}^{(\mathbf{l})}) \qquad (4.4.5)$$

where $\mathbf{z}^{(l)}$ is the output of the layer l, and $\mathbf{\Theta}^{(l+1)}$ is the parameter matrix and $a^{(l+1)}$ is the activation function of layer l+1 acting on $\mathbf{z}^{(l)}$.

Its inverse process, namely backward propagation[106, 107], is used to minimize the error of each layer in the learning process. In backward propagation, first, the error of the output layer $\boldsymbol{\delta}^{(\text{output layer})}$ is computed by subtracting the sample labels $\mathbf{y}$ from the corresponding predictions $\mathbf{a}^{(\text{output layer})}$:

$$\delta^{(output\ layer)} = a^{(output\ layer)}(\mathbf{z}^{(output\ layer)}) - \mathbf{y} \tag{4.4.6}$$

Next, $\boldsymbol{\delta}^{(\text{output layer})}$ is multiplied by the parameter matrix of the last hidden layer $\mathbf{\Theta}^{(\text{last hidden layer})}$ on the left and the gradient of its activation function $\mathbf{a'}^{(\text{last hidden layer})}$ to derive the error vector of the last hidden layer:

$$\delta^{(\text{last hidden layer})} = (\Theta^{(\text{last hidden layer})})^T \delta^{(\text{output layer})} a'^{(\text{last hidden layer})}(\mathbf{z}^{(\text{last hidden layer})}) \tag{4.4.7}$$

And so on and so forth until of error of all the layers are calculated. Finally, the error gradient for all the layers are calculated using equation 4.4.8: For a certain layer l, its error gradient matrix $\mathbf{\Delta}^{(l)}$ is computed by multiplying the error vector of the next layer($\boldsymbol{\delta}^{(l+1)}$ which is a column vector) with the transposed output of the present layer (($\mathbf{a}^{(l)})^T$ which is a row vector) divided by the number of samples m:

$$\Delta^{(l)} = \frac{1}{m}\delta^{(l+1)}(a^{(l)})^T \tag{4.4.8}$$

The gradients are then used to update the parameter matrix for all the layers.

However, a huge issue of a deeper network is the vanishing gradient[108]: in the forward propagation, the gradient signal declines to zero quickly when it goes deeper and deeper, while in the backward propagation, the weight matrices are multiplied at each step, making the gradient decrease exponentially to zero. To overcome this issue, we used the idea of and built a Residual Network (ResNet).[109]

In a ResNet, a shortcut allows the gradient to be directly propagated to the destination layers by skipping over several intermediate ones. Although the gradients still vanish on the main path, they don't vanish through propagations on the shortcut path due to the absence of the intermediate layers. In our ResNet model, we designed an identity block consisting of three convolution layers for the skip connection. Figure 4.4.3 shows its architecture:

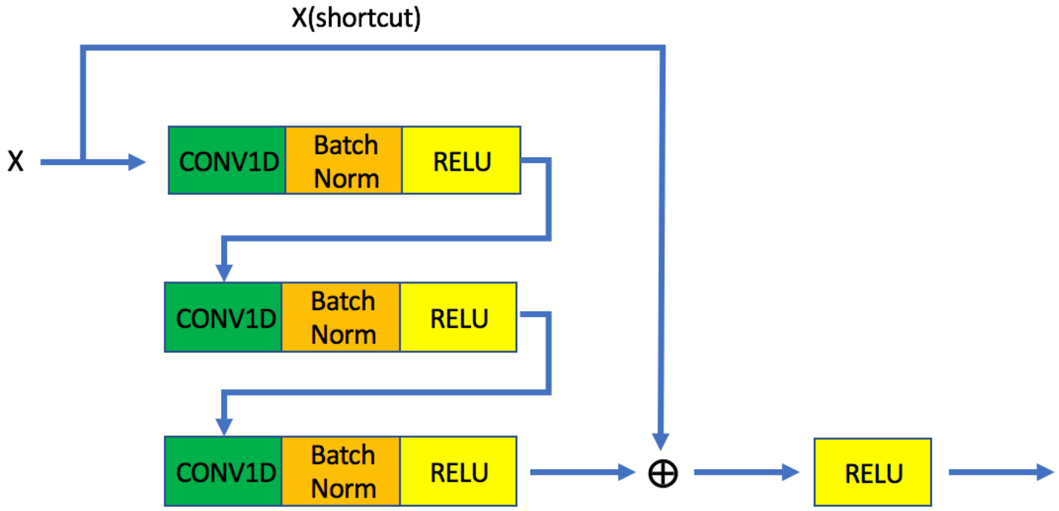

Figure 4.4.3 An identity block consisting of two paths.

The name "identity block" comes from the same dimension of the input (number of training samples × 600 features) and the output of the block. The block has two paths as shown in Figure 4.4.3. The upper path is the "shortcut path." The lower path is the "main path." In each layer of convolution, we also added the BatchNorm and the ReLU step. The first CONV1D has 32 filters of shape 1 and a stride of 1, the second CONV1D has 32 filters of shape 3 and a stride of 1, and the last one has 64 filters of shape 1 and a stride of 2. By use of these parameters, the input recovers its dimensionality of number of training samples × 600 features after the three convolution layers.

Our deeper CNN model consists of two parts. The first part is the previous simple CNN ended at the max pooling step which generated a matrix of 149×64 for each training sample (see Figure 4.4.2). The second part is the identity block of three convolution layer and a short cut, ending with the outputs of the same dimensionality. Finally the input goes into the flatten and fully connected layer and outputs the prediction. The resulting accuracies are listed below in Table 4.4.4:

training/testing accuracies	class 1	class 2	class 3
class 2	1.00/0.983	-	-
class 3	1.00/0.981	0.997/0.953	-
class 4	1.00/0.988	1.00/0.940	1.00/0.951

Table 4.4.4 Accuracies of binary classification using a deeper convolutional neural network with the identity block

By using a deeper convolutional network, the prediction results are further improved even though the input datasets are the same as used in the simple CNN model, which again only contains sequential information.

In order to test the robustness of this model, we included the nucleoporin (Nup) family as another class. Unlike the first four families, Nups are intrinsically disordered proteins (IDPs) that lack much secondary and tertiary structures[110-112]. The classification results are shown in Table 4.4.5:

training/testing accuracies	class 1	class 2	class 3	class4
Nup	1.0/0.970	0.958/0.727	0.972/0.911	0.998/0.946

Table 4.4.5 Accuracies of binary classifications between Nup and the first four families with the deep CNN model

These accuracies are lower than the results among the four families themselves, which may be attributed to different properties in sequence that IDPs have from the structured proteins. For example, many IDP sequences are of low complexity, meaning that they are over-represented by a few specific residues[113]. The lack of diversity of residues of the IDP sequences is undesired for classifications.

4.4.4 Model 3: LSTM network

Long short-term memory network (LSTM) is a special kind of recurrent neural network (RNN). A recurrent neural saves the output of a hidden layer and uses it along with the next sample as the inputs for making predictions as shown in Figure 4.4.4.

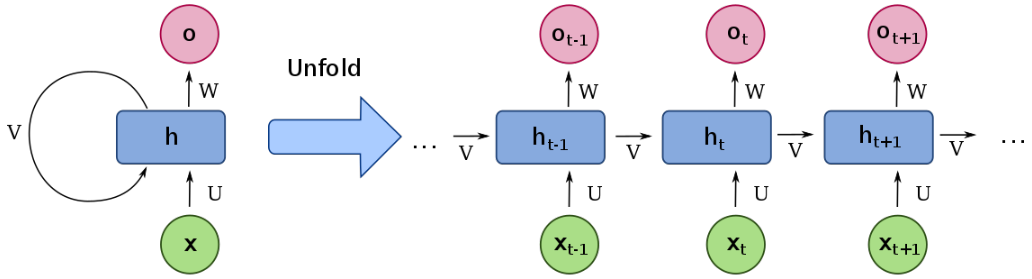

Figure 4.4.4 Architecture of a RNN. From https://en.wikipedia.org/wiki/Recurrent_neural_network

where $\mathbf{x_t}$ (number of features (m)×1) is the input vector of sample t, $\mathbf{U}$(number of neurons in the hidden layer (h)× number of features (m)) is the parameter matrix to derive $\mathbf{h_t}$ (h×1) which is the hidden state vector of dimension h. $\mathbf{h_t}$ is then multiplied by a parameter vector $\mathbf{W}$(m×h) to derive the output vector of the current sample $\mathbf{o_t}$ (m×1) and multiplied by a coefficient V as the weight and added to the next example

$\mathbf{h_{t+1}}$. σ_h and σ_o are the tanh activation function and sigmoid activation function for deriving $\mathbf{h_t}$ and $\mathbf{o_t}$, and $\mathbf{b_h}$ and $\mathbf{b_o}$ are corresponding bias vectors, respectively:

$$\mathbf{h}_t = \sigma_h(\mathbf{U}\mathbf{x}_t + \mathbf{b_h} + V\mathbf{h}_{t-1})$$

$$\mathbf{o_t} = \sigma_o(\mathbf{W}\mathbf{h}_t + \mathbf{b_o}) \tag{4.4.9}$$

In this way, a RNN connects the previous information to the present task by adding the hidden layer outputs of previous samples to the current input. The label information of all the previous samples is therefore accumulated to affect the prediction of the present sample. Thus RNN networks apply well to sequential information.

However, like all other multilayer neural networks, RNNs also suffer from the vanishing gradient problem. When the distance between the current point and the relevant sample in the past is long, RNN can hardly make use of the connection due to the vanishing gradient.[108]

Long Short Term Memory networks (LSTM) introduced by Hochreiter & Schmidhuber [14] was designed to solve long-term dependency problem. Figure 4.4.5 shows the repeating cell module of a LSTM, which has four layers.

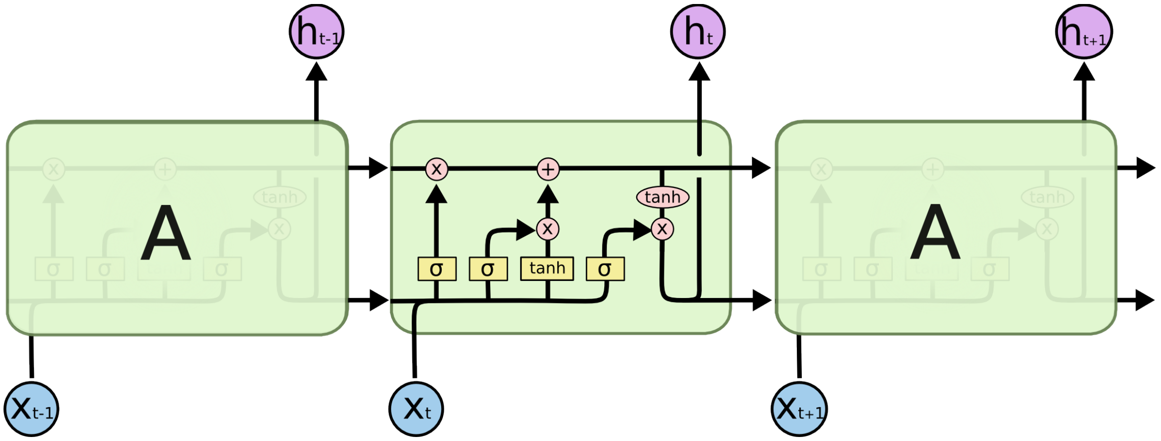

Figure 4.4.5 Structure of the repeating cell module in an LSTM. From http://colah.github.io/posts/2015-08-Understanding-LSTMs/

The first layer is the "forget gate layer", which controls the extent to which the information remains in the memory cell. The output vector $\mathbf{f_t}$(number of the neurons in the forget gate layer m×1) is given by:

$$\mathbf{f}_t = \sigma_g(\mathbf{W}_\mathbf{f}\mathbf{x}_\mathbf{t} + \mathbf{U}_\mathbf{f}\mathbf{h}_\mathbf{t-1} + \mathbf{b}_\mathbf{f}) \tag{4.4.10}$$

where $\mathbf{W_f}$ (m×d)and $\mathbf{U_f}$ (m×m) are the parameter matrices applying to the input vector $\mathbf{x_t}$(d×1) and the last hidden layer $\mathbf{h_{t-1}}$ (m×1). $\mathbf{b_f}$ (m×1)is the bias, σ_g is the sigmoid activation function.

The next step is to decide what new information to store in the memory cell. It has two paths. The first path is called the "input gate layer" $\mathbf{i_t}$(m×1)which controls the extent to which a new value flows into the cell. The second path outputs a candidate value $\tilde{\mathbf{C}}_\mathbf{t}$ (m×1) added to the cell state:

$$\begin{aligned} \mathbf{i}_\mathbf{t} &= \sigma_g(\mathbf{W}_\mathbf{i}\mathbf{x}_\mathbf{t} + \mathbf{U}_\mathbf{i}\mathbf{h}_\mathbf{t-1} + \mathbf{b}_\mathbf{i}) \\ \tilde{\mathbf{C}}_\mathbf{t} &= \sigma_C(\mathbf{W}_\mathbf{C}\mathbf{x}_\mathbf{t} + \mathbf{U}_\mathbf{c}\mathbf{h}_\mathbf{t-1} + \mathbf{b}_\mathbf{C}) \end{aligned} \tag{4.4.11}$$

where σ_C is the tanh activation function. $\mathbf{W_i}$ (m×d), $\mathbf{W_c}$ (m×d), and $\mathbf{U_i}$ (m×m), $\mathbf{U_c}$(m×m) are the parameter matrices that apply to the **x** and **h** as mentioned above. $\mathbf{b_i}$ (m×1)and $\mathbf{b_c}$ (m×1)are the corresponding bias vectors for $\mathbf{i_t}$ and $\tilde{\mathbf{C}}_\mathbf{t}$.

Thus the new state of the cell is updated from $\mathbf{C_{t-1}}$(m×1) to $\mathbf{C_t}$ (m×1) by

$$\mathbf{C_t} = \mathbf{f_t} \circ \mathbf{C_{t\text{-}1}} + \mathbf{i_t} \circ \tilde{\mathbf{C}}_\mathbf{t} \tag{4.4.12}$$

Note that the operator $\circ$ denotes the entry-wise product.

Finally, the output vector of the cells from the last layer “output gate”, $\mathbf{o_t}$(m×1), is given by

$$\begin{aligned} \mathbf{o_t} &= \sigma_g(\mathbf{W_o}\mathbf{x_t} + \mathbf{U_o}\mathbf{h_{t\text{-}1}} + \mathbf{b_o}) \\ \mathbf{h_t} &= \mathbf{o_t} \circ \sigma_C(\mathbf{C_t}) \end{aligned} \tag{4.4.13}$$

with $\mathbf{W_o}$ (m×d), $\mathbf{U_o}$ (m×m), and $\mathbf{b_o}$ (m×1). The hidden layer vector $\mathbf{h_t}$ (m×1) that goes to the next sample is also obtained by equation 4.4.13.

We used the Keras.layers.LSTM package to implement the LSTM network with the inputs of the four classes mentioned in the Section 4.2 (class 5 to 8: Helicase, Transmembrane receptor, Amino acid kinase, ATPase) from the PDB website[98] obtained by four different methods:

1) sequential data representing the primary structure information,

2) add in structural data recording the residue types that enter the neighborhood of radius 8Å of the first 100 residues,

3) add in structural data and recording the number of occurrences of the neighbors with neighborhood radius: 6Å and

4) add in structural data and recording the number of occurrences of the neighbors with neighborhood radius: 12Å.

The training set contains 600 samples in total while the testing set contains 200 samples. The results obtained from the different input sources are shown in Tables 4.4.6-9 in the order as above:

training/testing accuracies	class 5	class 6	class 7
class 6	1.00/0.750	-	-
class 7	1.00/0.815	0.982/0.820	-
class 8	0.958/0.705	1.0/0.873	1.00/0.850

Table 4.4.6 Method 1: Accuracies of LSTM using the sequential representation

training/testing accuracies	class 5	class 6	class 7
class 6	1.00/0.830	-	-
class 7	1.00/0.918	0.987/0.828	-
class 8	0.963/0.730	1.00/0.905	1.00/0.893

Table 4.4.7 Method 2: Accuracies of LSTM adding in the first structural representation with threshold = 8Å

training/testing accuracies	class 5	class 6	class 7
class 6	1.00/0.886	-	-
class 7	1.00/0.872	0.991/0.876	-
class 8	0.965/0.789	0.997/0.897	1.00/0.898

Table 4.4.8 Method 3: Accuracies of LSTM adding in the second structural representation with threshold = 6Å

training/testing accuracies	class 5	class 6	class 7
class 6	0.996/0.880	-	-
class 7	0.997/0.908	0.989/0.903	-
class 8	0.974/0.790	0.993/0.901	0.983/0.892

Table 4.4.9 Method 4: Accuracies of LSTM adding in the second structural representation with threshold = 12Å

The results obtained from these four methods show that the first way of using only the sequential information gives the lowest prediction accuracies. The accuracies increase when the structural information is introduced. Then the accuracies are further increased when the occurrences of the neighbor type are counted as in Tables 4.4.8 and 4.4.9. The resulting accuracies from the 6Å threshold in Table 4.4.8 show a heavier overfitting compared to the those from the 12Å threshold in Table 4.4.9.

4.5 Conclusion and Remarks

To study the protein classification problem, we implement three models: model 1) a vanilla network, model 2) a CNN and model 3) an LSTM. For the CNN model, we tried a simple network with one convolution layer and a deeper one with three convolution layers and a shortcut path. The prediction accuracies show that a deep CNN solves the overfitting problem of the vanilla network by increasing the test accuracies efficiently. For the LSTM model, we compared two ways to extract input data from the PDB files. The first way gives the sequential representation of proteins, while the second way includes some structural information. Namely, information about the residue types in the neighborhood of each residue.

The results show that the introduction of such structural information significantly increases the testing accuracies for all the pairs in comparison with the training set of the same sample size which has only the sequential information in it.

However, we note that the testing accuracy obtained by using the structural information in model 3 is still much lower than the results in model 2. This is attributed to the smaller number of the training samples (400 per family) that we have from the PDB structure files for model 3 compared to the training sample size (7500 per family) for model 2. Note that 400 samples per family are too few to run model 2 since it starts with 600 features (type of the first 600 residues) that is even greater than 400 which leads to significantly lowered accuracies for both training and testing set and a dramatic overfitting problem (see Table 4.5.1):

training/testing accuracies	class 1	class 2	class 3
class 2	0.828/0.613	-	-
class 3	0.958/0.884	0.890/0.764	-
class 4	0.863/0.814	0.865/0.729	0.833/0.704

Table 4.5.1 Accuracies of binary classification using the deeper convolutional neural network in Model 2 with 400 sequence for each class

Thus the combination of sequence and structure does provide much more information than sequence alone.

5 Conclusion and Outlook

5.1 Intercalation Outlook

We investigated the Molecular Dynamics of S0/S1 YOPRO in different environment and revealed the relationship between the enhanced fluorescent intensity of intercalated YOPRO in dsDNA and its configuration distribution. YOPRO is the monomer of the oxazole yellow homodimer YOYO[71] (Figure 5.1.1). The use of dimeric dyes greatly enhances the fluorescence intensity due to both aromatic systems intercalating and accordingly dimers are typically used in assays.

According to the studies by Franci Johansen et al., YOYO strongly favors bis-intercalating to the (5'-CTAG-3')$_2$ binding sites in oligonucleotide d(CGCTAGCG)$_2$ and to the (5'-CCGG-3')$_2$ binding sites in oligonucleotide d(CGCCGGCG)$_2$ [64]. Examination of our dsDNA structure shows that it has two d1(CT):d2(GA) sites and one d1(GA):d2(CT) sites as well as two d1(CC):d2(GG) sites. All these five sites are suitable for intercalation. In our study, we only picked d1(5'-C4T5-3'):d2(3'-G38A37-5') for intercalation as described in Section 3.2. Other binding sites can be investigated as well to verify and generalize our observations and conclusions in the future.

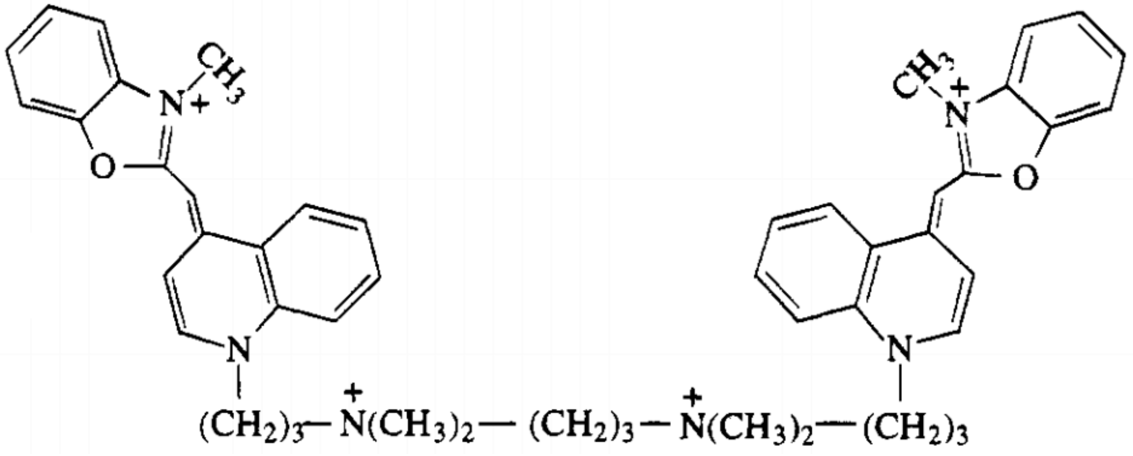

Figure 5.1.1 Chemical structure of YOYO (1,1'-(4,4,8,8- tetramethyl-4,8-diazaundecamethylene)bis[4-[3-methyl-benzo- 1 OX- azol-2-yl]methylidene]-1,4-dihydroquinolinium])[71].

We used constraint MD in water to gradually separate the neighbor base pairs on the binding site to provide intercalation space for YOPRO. Then the YOPRO at its planar conformations was manually docked into the binding site. In drug design and protein design projects, researchers also use docking softwares such as 1-Click Docking[114], Blaster[115] and DOCK[116], which automatically optimize the binding orientation with the binding affinities provided. These methodologies will be of particular use for intercalation of dimeric dyes.

5.2 Molecular Dynamics of MiR-155

We propose to carry out MD simulations of miR-155 (Figure 5.2.1), a microRNA (miRNA) present in humans encoded by the MIR155 host gene. microRNAs play a key role in the regulation of gene expression[117, 118]. Human miRNAs silence messenger RNA (mRNA) at the post-transcription stage by binding through complementary matches with a relatively small number of the bases (as few as 6-8) in the messenger RNA[119-121]. This binding prevents translation of the mRNA into a protein at the ribosome.

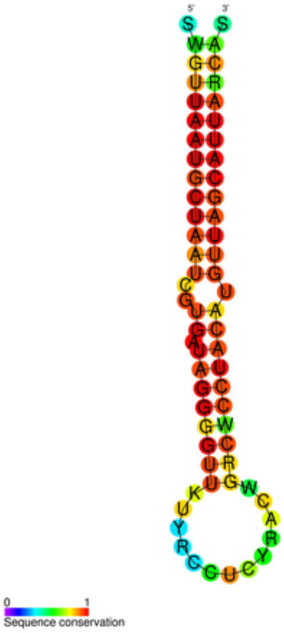

Figure 5.2.1 miR-155 secondary structure and sequence conservation. https://commons.wikimedia.org/w/index.php?curid=24565688

For the molecular dynamics studies of miR-155, we will use the AMBER force field for RNA, drawing on a recent revision of the RNA dihedral parameters[18]. We will examine the motion of a double-stranded species in which miR-155 is paired with a complementary strand of RNA. We also propose to simulate RNA/RNA interactions between the miR-155 and RNA segments that contain a sequence of 6 to 8 complementary bases, but then have base mismatches at other positions.

We will take advantage of the AMBER tool nucleic acid builder (nab)[18], which facilitates construction of RNA/RNA duplexes. Given the sequence for miR-155 above and its complement, nab will properly pair the complementary bases, thus providing the starting geometry for the simulations of the duplexes. The behavior of the partially matched double RNA strands will be challenging to determine, because of the increased dynamical degrees of freedom, but MD studies should still be feasible.

Fluorescent labeling has been established for miRNA duplexes[122], although the stable binding coordinates for the dye and the RNA duplex is not known. This must be determined prior to MD simulations. Fluorescence microscopy studies suggest that a surface binding mode is important for miRNA duplexes[123], although an intercalation mode should also be possible. In our work on DNA/YOPRO, we have found a surface-binding mode (see Section 3.3), and this can be used as a starting point to locate dye-RNA duplex surface binding coordinates. Molecular dynamics modeling of the motion of the complex can then proceed. We will also examine whether any increase in fluorescence intensity is predicted for an oxazole dye label on single-stranded miRNA. Single-stranded miRNA is often imaged after complexing with a luciferase reporter vector[124-126] rather than a fluorescent dye.

5.2.1

APPENDICES

APPENDIX A Fitting Protocol from IpolQ Method

Input file: “**d1__d2_kcal.txt**” which is subtracted surface with d1 fixed and d2 varied. The content has three columns corresponding to d1, d2 and Q-V in the units of kcal. Note each two adjacent fixed d1 values by the string “9999 9999 9999” which is the separation line.

1. run **d1_*make_slice.cpp***, get a bunch of “**d1_at_X.txt**”, where “X” is the fixed d1 location. Collect all “X” values to make the file “**fix_d1_namelist.txt**” which has all fixed D1 values.

2. run ***d1_amber_fit.m***, get the output “**amber_YOPRO_s0_4cos.txt**”, which is a diary file that contains the fitted force constant parameters for all the terms in eq. 1.4.1 at different fixed D1 values.

3. run ***d1_make_gaffpieces.cpp***, get “**fix_d1_fc_slice.txt**” which is extraction from “**amber_YOPRO_s0_4cos.txt**” that contains all force constant parameter.

Now do the same thing to D2:

Input file : “**d1__d2_kcal_gnup.txt**” is a copy of “**d1__d2_kcal.txt**” with all separation lines “9999 9999 9999” deleted.

4. run ***transpose.cpp*** to transpose **d1__d2_kcal_gnup.txt** into “**d2__d1_kcal.txt**”. Add the separation lines to separate adjacent d2 values.

5. run **d2_make_slice.cpp**, get a bunch of “**d2_at_X.txt**”, where “X” is the fixed d2 location. Collect all “X” values to make the file “**fix_d2_namelist.txt**”.

6. run ***d2_make_gaffpieces.cpp***, get “**fix_d2_fc_slice.txt**”.

Now make AMBER frcmod files for umbrella samplings:

Input file: “**YOPRO.frcmod**” is a frcmod template file.

7. run ***creatfrcmod.cpp*** to obtain all frcmod files for the umbrella samplings.

APPENDIX B C++ program implementing the steepest descent algorithm

```
#include<iostream>
#include<string>
#include<vector>
#include<fstream>
#include<utility>
using namespace std;
int main()
{
   ifstream input;
   string in_name = "";
   cout<<"input a file"<<endl;
   cin>>in_name;
   input.open(in_name);
   if(!input.is_open())
   {
      cout<<"No such file!"<<endl;
      return 0;
   }

   double xbsize,ybsize=0;
   cout<<"input xbin size, ybin size:"<<endl;
   cin>>xbsize>>ybsize;
```

```
double xmax,xmin,ymax,ymin =0;

cout<<"input xmin,xmax,ymin,ymax"<<endl;

cin>>xmin>>xmax>>ymin>>ymax;


int xbnum = 1+(xmax-xmin)/xbsize;

int ybnum = 1+(ymax-ymin)/ybsize;


double** PMF = new double* [xbnum];

for(int i=0;i<xbnum;i++)

{

   PMF[i] = new double[ybnum];

}


string skip="";

getline(input,skip);


for(int i=0;i<xbnum;i++)

{

   for(int j=0;j<ybnum;j++)

   {

      input>>skip>>skip>>PMF[i][j]>>skip;

      //cout<<PMF[i][j]<<endl;
```

```
  }
}

double x_st,y_st=0;
cout<<"starting point:"<<endl;
cin>>x_st>>y_st;
int x_index = (x_st-xmin)/xbsize;
int y_index = (y_st-ymin)/ybsize;
double PMF_st = PMF[x_index][y_index];
ofstream output(in_name+"result.txt");
vector<pair<int,int>> x_y;
x_y.push_back(make_pair(x_index,y_index));
cout<<xmin+x_index*xbsize<<" "<<ymin+y_index*ybsize<<" "<<PMF_st<<endl;
output<<xmin+x_index*xbsize<<" "<<ymin+y_index*ybsize<<" "<<PMF_st<<endl;

double neighbor_up = PMF[x_index][y_index];
double neighbor_down = PMF[x_index][y_index];
double neighbor_right = PMF[x_index][y_index];
//double PMF_next = PMF[x_index][y_index];

while(x_index<=xbnum-1 && x_index>=0 && y_index<=ybnum-1 && y_index>=0)
{
  if(y_index<ybnum-1)
```

```
    neighbor_up = PMF[x_index][y_index+1];

    if(y_index>0)
    neighbor_down = PMF[x_index][y_index-1];

    if(x_index<xbnum-1)
    neighbor_right = PMF[x_index+1][y_index];
    else
    break;

    double PMF_old =PMF[x_index][y_index];
    double PMF_next = PMF[x_index][y_index];

    if(neighbor_up<=neighbor_down && neighbor_up<=neighbor_right /*&&
neighbor_up<=PMF_old*/)
    {
      y_index++;
      if(x_y.size()>2)
      {
        if(y_index==x_y[x_y.size()-2].second && x_index == x_y[x_y.size()-2].first)
//prevent stepping back and forth
        {
          y_index--;
          if(neighbor_right<=neighbor_down)
          {
            PMF_next = neighbor_right;
            x_index++;
          }
          else
          {
            PMF_next = neighbor_down;
            y_index--;
          }
          if(PMF_next==PMF_old)
          {
            cout<<"done"<<endl;
            break;
          }
          x_y.push_back(make_pair(x_index,y_index));
          cout<<xmin+x_index*xbsize<<" "<<ymin+y_index*ybsize<<"
"<<PMF_next<<endl;
          output<<xmin+x_index*xbsize<<" "<<ymin+y_index*ybsize<<"
"<<PMF_next<<endl;
          continue;
        }
      }
```

```
          PMF_next = neighbor_up;
        }
        if(neighbor_down<=neighbor_up && neighbor_down<=neighbor_right /*&&
neighbor_down<=PMF_old*/)
        {
          y_index--;
          if(x_y.size()>2)
          {
            if(y_index==x_y[x_y.size()-2].second && x_index == x_y[x_y.size()-2].first)
//prevent stepping back and forth
            {
              y_index++;
              if(neighbor_right<=neighbor_up)
              {
                PMF_next = neighbor_right;
                x_index++;
              }
              else
              {
                PMF_next = neighbor_up;
                y_index++;
              }
              if(PMF_next==PMF_old)
```

```
                {
                    cout<<"done"<<endl;
                    break;
                }
                x_y.push_back(make_pair(x_index,y_index));
                cout<<xmin+x_index*xbsize<<" "<<ymin+y_index*ybsize<<"
"<<PMF_next<<endl;
                output<<xmin+x_index*xbsize<<" "<<ymin+y_index*ybsize<<"
"<<PMF_next<<endl;
                continue;
            }
        }
        PMF_next = neighbor_down;
    }
    if(neighbor_right<=neighbor_up && neighbor_right<=neighbor_down /*&&
neighbor_right<=PMF_old*/)
    {
        x_index++;
        PMF_next = neighbor_right;
    }
    if(PMF_next==PMF_old)
    {
        cout<<"done"<<endl;
```

```
      break;
    }
    x_y.push_back(make_pair(x_index,y_index));
    cout<<xmin+x_index*xbsize<<" "<<ymin+y_index*ybsize<<"
"<<PMF_next<<endl;
    output<<xmin+x_index*xbsize<<" "<<ymin+y_index*ybsize<<"
"<<PMF_next<<endl;
  }
}
```

REFERENCES

REFERENCES

1. Rye, H. S.; Yue, S.; Wemmer, D. E.; Quesada, M. A.; Haugland, R. P.; Mathies, R. A.; Glazer, A. N. Stable Fluorescent Complexes of Double-Stranded DNA with Bis-Intercalating Asymmetric Cyanine Dyes - Properties and Applications. *Nucleic Acids Res* **1992,** *20*, 2803-2812.

2. Selvin, P. Science innovation '92: the San Francisco sequel. *Science* **1992,** *257*, 885-6.

3. Glazer, A. N.; Rye, H. S. Stable dye-DNA intercalation complexes as reagents for high-sensitivity fluorescence detection. *Nature* **1992,** *359*, 859-61.

4. Rye, H. S.; Yue, S.; Quesada, M. A.; Haugland, R. P.; Mathies, R. A.; Glazer, A. N. Picogram detection of stable dye-DNA intercalation complexes with two-color laser-excited confocal fluorescence gel scanner. *Methods Enzymol* **1993,** *217*, 414-31.

5. Bianco, P. R.; Brewer, L. R.; Corzett, M.; Balhorn, R.; Yeh, Y.; Kowalczykowski, S. C.; Baskin, R. J. Processive translocation and DNA unwinding by individual RecBCD enzyme molecules. *Nature* **2001,** *409*, 374-8.

6. Leuba, S. H.; Anand, S. P.; Harp, J. M.; Khan, S. A. Expedient placement of two fluorescent dyes for investigating dynamic DNA protein interactions in real time. *Chromosome Res* **2008,** *16*, 451-67.

7. Armitage, B. A. Cyanine dye-DNA interactions: Intercalation, groove binding, and aggregation. *Top Curr Chem* **2005,** *253*, 55-76.

8. Hirons, G. T.; Fawcett, J. J.; Crissman, H. A. TOTO and YOYO: New Very Bright Fluorochromes for DNA Content Analyses by Flow-Cytometry. *Cytometry* **1994,** *15*, 129-140.

9. Figeys, D.; Arriaga, E.; Renborg, A.; Dovichi, N. J. Use of the Fluorescent Intercalating Dyes Popo-3, Yoyo-3 and Yoyo-1 for Ultrasensitive Detection of Double-Stranded DNA Separated by Capillary Electrophoresis with Hydroxypropylmethyl Cellulose and Non-Cross-Linked Polyacrylamide. *J Chromatogr A* **1994,** *669*, 205-216.

10. Netzel, T. L.; Nafisi, K.; Zhao, M.; Lenhard, J. R.; Johnson, I. Base-content dependence of emission enhancements, quantum yields, and lifetimes for cyanine dyes bound to double-strand DNA: Photophysical properties of monomeric and bichromophoric DNA stains. *J Phys Chem-Us* **1995,** *99*, 17936-17947.

11. Cerutti, D. S.; Rice, J. E.; Swope, W. C.; Case, D. A. Derivation of Fixed Partial Charges for Amino Acids Accommodating a Specific Water Model and Implicit Polarization. *J Phys Chem B* **2013,** *117*, 2328-2338.

12. Bateman, A.; Coin, L.; Durbin, R.; Finn, R. D.; Hollich, V.; Griffiths‐Jones, S.; Khanna, A.; Marshall, M.; Moxon, S.; Sonnhammer, E. L. The Pfam protein families database. *Nucleic Acids Res* **2004,** *32*, D138-D141.

13. van Gerven, M.; Bohte, S. Editorial: Artificial Neural Networks as Models of Neural Information Processing. *Front Comput Neurosc* **2017,** *11*.

14. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput* **1997,** *9*, 1735-1780.

15. Alder, B. J.; Wainwright, T. E. Molecular Motions. *Sci Am* **1959,** *201*, 113-&.

16. Alder, B. J.; Wainwright, T. E. Studies in Molecular Dynamics .1. General Method. *J Chem Phys* **1959,** *31*, 459-466.

17. Allen, M. P.; Tildesley, D. J. *Computer simulation of liquids*. Oxford university press: 2017.

18. Case, D.; Darden, T.; Cheatham III, T.; Simmerling, C.; Wang, J.; Duke, R.; Luo, R.; Walker, R.; Zhang, W.; Merz, K. AMBER 12; University of California: San Francisco, 2012. *There is no corresponding record for this reference* **2010**, 1-826.

19. Mackerell, A. D.; Feig, M.; Brooks, C. L. Extending the treatment of backbone energetics in protein force fields: Limitations of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulations. *J Comput Chem* **2004,** *25*, 1400-1415.

20. van Gunsteren, W. F.; Billeter, S. R.; Eising, A. A.; Hünenberger, P. H.; Krüger, P.; Mark, A. E.; Scott, W. R.; Tironi, I. G. Biomolecular simulation: the {GROMOS96} manual and user guide. **1996**.

21. Hockney, R.; Eastwood, J. Computer Simulation Using Particles 1988. *Adam Hilger* **1988**, 120-128.

22. Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. G. A Smooth Particle Mesh Ewald Method. *J Chem Phys* **1995,** *103*, 8577-8593.

23. Smit, B. *Understanding molecular simulation: from algorithms to applications*. Academic Press: 1996.

24. Berendsen, H. J. C.; Postma, J. P. M.; Vangunsteren, W. F.; Dinola, A.; Haak, J. R. Molecular-Dynamics with Coupling to an External Bath. *J Chem Phys* **1984,** *81*, 3684-3690.

25. Andersen, H. C. Molecular-Dynamics Simulations at Constant Pressure and/or Temperature. *J Chem Phys* **1980,** *72*, 2384-2393.

26. Karasawa, N.; Goddard, W. A. Acceleration of Convergence for Lattice Sums. *J Phys Chem-Us* **1989,** *93*, 7320-7327.

27. Kumar, S.; Bouzida, D.; Swendsen, R. H.; Kollman, P. A.; Rosenberg, J. M. The Weighted Histogram Analysis Method for Free-Energy Calculations on Biomolecules .1. The Method. *J Comput Chem* **1992,** *13*, 1011-1021.

28. Souaille, M.; Roux, B. Extension to the weighted histogram analysis method: combining umbrella sampling with free energy calculations. *Comput Phys Commun* **2001,** *135*, 40-57.

29. McQuarrie, D. Statistical mechanics. *Sausalito, Calif.: University Science Books* **2004,** *12*, 641.

30. Kirkwood, J. G. Statistical mechanics of fluid mixtures. *J Chem Phys* **1935,** *3*, 300-313.

31. Cukier, R. I. Variance of a Potential of Mean Force Obtained Using the Weighted Histogram Analysis Method. *J Phys Chem B* **2013,** *117*, 14785-14796.

32. Roux, B. The Calculation of the Potential of Mean Force Using Computer-Simulations. *Comput Phys Commun* **1995,** *91*, 275-282.

33. Haverkort, F.; Stradomska, A.; de Vries, A. H.; Knoester, J. Investigating the Structure of Aggregates of an Amphiphilic Cyanine Dye with Molecular Dynamics Simulations. *J Phys Chem B* **2013,** *117*, 5857-5867.

34. Neese, F. The ORCA program system. *Wires Comput Mol Sci* **2012,** *2*, 73-78.

35. Bayly, C. I.; Cieplak, P.; Cornell, W. D.; Kollman, P. A. A Well-Behaved Electrostatic Potential Based Method Using Charge Restraints for Deriving Atomic Charges - the Resp Model. *J Phys Chem-Us* **1993,** *97*, 10269-10280.

36. Turro, N. J.; Ramamurthy, V.; Scaiano, J. C. Modern Molecular Photochemistry of Organic Molecules. *Photochem Photobiol* **2012,** *88*, 1033-1033.

37. Hirons, G. T.; Fawcett, J. J.; Crissman, H. A. Toto and Yoyo - New Very Bright Fluorochromes for DNA Content Analyses by Flow-Cytometry. *Cytometry* **1994,** *15*, 129-140.

38. Carlsson, C.; Larsson, A.; Jonsson, M.; Albinsson, B.; Norden, B. Optical and Photophysical Properties of the Oxazole Yellow DNA Probes Yo and Yoyo. *J Phys Chem-Us* **1994,** *98*, 10313-10321.

39. Cao, J. F.; Hu, C.; Liu, F.; Sun, W.; Fan, J. L.; Song, F. L.; Sun, S. G.; Peng, X. J. Mechanism and Nature of the Different Viscosity Sensitivities of Hemicyanine Dyes with Various Heterocycles. *Chemphyschem* **2013,** *14*, 1601-1608.

40. Kummer, A. D.; Kompa, C.; Niwa, H.; Hirano, T.; Kojima, S.; Michel-Beyerle, M. E. Viscosity-dependent fluorescence decay of the GFP chromophore in solution due to fast internal conversion. *J Phys Chem B* **2002,** *106*, 7554-7559.

41. Upadhyayula, S.; Nunez, V.; Espinoza, E. M.; Larsen, J. M.; Bao, D.; Shi, D. W.; Mac, J. T.; Anvari, B.; Vullev, V. I. Photoinduced dynamics of a cyanine dye: parallel pathways of non-radiative deactivation involving multiple excited-state twisted transients. *Chem Sci* **2015,** *6*, 3269-3269.

42. Anni, M.; Della Sala, F.; Raganato, M. F.; Fabiano, E.; Lattante, S.; Cingolani, R.; Gigli, G.; Barbarella, G.; Favaretto, L.; Gorling, A. Nonradiative relaxation in Thiophene-S,S-dioxide derivatives: The role of the environment. *J Phys Chem B* **2005,** *109*, 6004-6011.

43. Sharafy, S.; Muszkat, K. A. Viscosity Dependence of Fluorescence Quantum Yields. *J Am Chem Soc* **1971,** *93*, 4119-&.

44. Saltiel, J.; Zafiriou, O. C.; Megarity, E. D.; Lamola, A. A. Tests of Singlet Mechanism for Cis-Trans Photoisomerization of Stilbenes. *J Am Chem Soc* **1968,** *90*, 4759-&.

45. Kramers, H. A. Brownian motion in a field of force and the diffusion model of chemical reactions. *Physica* **1940,** *7*, 284-304.

46. Hanggi, P.; Talkner, P.; Borkovec, M. Reaction-Rate Theory - 50 Years after Kramers. *Rev Mod Phys* **1990,** *62*, 251-341.

47. Risken, H. Fokker-planck equation. In *The Fokker-Planck Equation*, Springer: 1996; pp 63-95.

48. Sanchez-Galvez, A.; Hunt, P.; Robb, M. A.; Olivucci, M.; Vreven, T.; Schlegel, H. B. Ultrafast radiationless deactivation of organic dyes: Evidence for a two-state two-mode pathway in polymethine cyanines. *J Am Chem Soc* **2000,** *122*, 2911-2924.

49. Rettig, W. Charge Separation in Excited-States of Decoupled Systems - Tict Compounds and Implications Regarding the Development of New Laser-Dyes and the Primary Processes of Vision and Photosynthesis. *Angew Chem Int Edit* **1986,** *25*, 971-988.

50. Sczepan, M.; Rettig, W.; Bricks, Y. L.; Slominski, Y. L.; Tolmachev, A. I. Unsymmetric cyanines: chemical rigidization and photophysical properties. *J Photoch Photobio A* **1999,** *124*, 75-84.

51. Gao, A. H.; Zhang, P. Y.; Zhao, M. Y.; Liu, J. Y. Photoisomerization mechanism of 1,1 '-dimethyl-2,2 '-pyridocyanine in the gas phase and in solution. *Spectrochim Acta A* **2015,** *136*, 1157-1166.

52. Tomasi, J.; Mennucci, B.; Cammi, R. Quantum mechanical continuum solvation models. *Chem Rev* **2005,** *105*, 2999-3093.

53. http://www.ebi.ac.uk/chebi/.

54. Frisch, M. J. Gaussian 03.

55. Singh, U. C.; Kollman, P. A. An Approach to Computing Electrostatic Charges for Molecules. *J Comput Chem* **1984,** *5*, 129-145.

56. D.A. Case, T. A. D., T.E. Cheatham, III, C.L. Simmerling, J. Wang, R.E. Duke, R.; Luo, R. C. W., W. Zhang, K.M. Merz, B. Roberts, S. Hayik, A. Roitberg, G. Seabra,; J. Swails, A. W. G., I. Kolossváry, K.F. Wong, F. Paesani, J. Vanicek, R.M. Wolf, J. Liu,; X. Wu, S. R. B., T. Steinbrecher, H. Gohlke, Q. Cai, X. Ye, J. Wang, M.-J. Hsieh, G.; Cui, D. R. R., D.H. Mathews, M.G. Seetin, R. Salomon-Ferrer, C. Sagui, V. Babin, T.; Luchko, S. G., A. Kovalenko, and P.A. Kollman AMBER 12. **2012.**

57. Ryckaert, J. P.; Ciccotti, G.; Berendsen, H. J. C. Numerical-Integration of Cartesian Equations of Motion of a System with Constraints - Molecular-Dynamics of N-Alkanes. *J. Comput. Phys.* **1977,** *23*, 327-341.

58. Berendsen, H. H. C.; Postma, J. P. M.; Gunsteren, W. F.; DiNola, A.; Haak, J. R. Molecular dynamics with coupling to an external bath. *J. Chem. Phys.* **1984,** *81*, 3684-3690.

59. Ellenberger, T. E.; Brandl, C. J.; Struhl, K.; Harrison, S. C. The Gcn4 Basic Region Leucine Zipper Binds DNA as a Dimer of Uninterrupted Alpha-Helices - Crystal-Structure of the Protein-DNA Complex. *Cell* **1992,** *71*, 1223-1237.

60. Schrodinger, LLC, The PyMOL Molecular Graphics System, Version 1.3r1. 2010.

61. Grossfield, A. WHAM: the weighted histogram analysis method. http://membrane*. urmc. rochester. edu/content/wham* **2012.**

62. Souaille, M.; Roux, B. Extension to the weighted histogram analysis method: combining umbrella sampling with free energy calculations. *Comput. Phys. Commun.* **2001,** *135*, 40-57.

63. Humphrey, W.; Dalke, A.; Schulten, K. VMD: Visual molecular dynamics. *Journal of Molecular Graphics & Modelling* **1996,** *14*, 33-38.

64. Johansen, F.; Jacobsen, J. P. H-1 NMR studies of the bis-intercalation of a homodimeric oxazole yellow dye in DNA oligonucleotides. *J Biomol Struct Dyn* **1998,** *16*, 205-222.

65. http://www.ebi.ac.uk/chebi/.

66. Allen, M. P.; Tildesley, D. J. *Computer Simulation of Liquids*. Clarendon Press: Oxford, 1987.

67. Frenkel, D.; Smit, B. *Understanding Molecular Simulation: From Algorithms to Applications*. Academic: San Diego, 1996.

68. Yesudas, K. Cationic cyanine dyes: impact of symmetry-breaking on optical absorption and third-order polarizabilities. *Phys. Chem. Chem. Phys.* **2013,** *15*, 19465-19477.

69. Zhang, X.; Wang, L.; Zhai, G.; Wen, Z.; Zhang, Z. The absorption, emission spectra as well as ground and excited states calculations of some dimethine cyanine dyes. *J. Mol. Struc-Theochem.* **2009,** *906*, 50-55.

70. Kuhn, H. A Quantum-Mechanical Theory of Light Absorption of Organic Dyes and Similar Compounds. *J. Chem. Phys.* **1949,** *17*, 1198-1212.

71. Larsson, A.; Carlsson, C.; Jonsson, M.; Albinsson, B. Characterization of the Binding of the Fluorescent Dyes Yo and Yoyo to DNA by Polarized-Light Spectroscopy. *J Am Chem Soc* **1994,** *116*, 8459-8465.

72. Rye, H. S.; Dabora, J. M.; Quesada, M. A.; Mathies, R. A.; Glazer, A. N. Fluorometric Assay Using Dimeric Dyes for Double-Stranded and Single-Stranded-DNA and Rna with Picogram Sensitivity. *Anal Biochem* **1993,** *208*, 144-150.

73. Jin, C.; Cerutti, D.; Cukier, R. I. Molecular Dynamics of Oxazole Yellow Dye in its Ground and First Excited Electronic States in Solution and when Intercalated in dsDNA. *J Phys Chem B* **2017,** *121*, 10242-10248.

74. Gao, A.; Zhang, P.; Zhao, M.; Liu, J. Photoisomerization mechanism of 1,1'-dimethyl-2,2'-pyridocyanine in the gas phase and in solution. *Spectrochim Acta A Mol Biomol Spectrosc* **2015,** *136 Pt B*, 1157-66.

75. Improta, R.; Santoro, F. A theoretical study on the factors influencing cyanine photo isomerization: The case of thiacyanine in gas phase and in methanol. *Journal of Chemical Theory and Computation* **2005,** *1*, 215-229.

76. Murphy, S.; Schuster, G. B. Electronic Relaxation in a Series of Cyanine Dyes - Evidence for Electronic and Steric Control of the Rotational Rate. *Journal of Physical Chemistry* **1995,** *99*, 8516-8518.

77. Sundstrom, V.; Vangrondelle, R.; Bergstrom, H.; Akesson, E.; Gillbro, T. Excitation-Energy Transport in the Bacteriochlorophyll Antenna Systems of Rhodospirillum-Rubrum and Rhodobacter-Sphaeroides, Studied by Low-Intensity Picosecond Absorption-Spectroscopy. *Biochimica Et Biophysica Acta* **1986,** *851*, 431-446.

78. Dietzek, B.; Bruggemann, B.; Pascher, T.; Yartsev, A. Mechanisms of molecular response in the optimal control of photoisomerization. *Physical Review Letters* **2006,** *97*.

79. Dietzek, B.; Tarnovsky, A. N.; Yartsev, A. Visualizing overdamped wavepacket motion: Excited-state isomerization of pseudocyanine in viscous solvents. *Chemical Physics* **2009,** *357*, 54-62.

80. Furstenberg, A.; Julliard, M. D.; Deligeorgiev, T. G.; Gadjev, N. I.; Vasilev, A. A.; Vauthey, E. Ultrafast excited-state dynamics of DNA fluorescent intercalators: New insight into the fluorescence enhancement mechanism. *Journal of the American Chemical Society* **2006,** *128*, 7661-7669.

81. Furstenberg, A.; Vauthey, E. Ultrafast excited-state dynamics of oxazole yellow DNA intercalators. *Journal of Physical Chemistry B* **2007,** *111*, 12610-12620.

82. Sundstrom, V.; Gillbro, T. Viscosity Dependent Radiationless Relaxation Rate of Cyanine Dyes - a Picosecond Laser Spectroscopy Study. *Chemical Physics* **1981,** *61*, 257-269.

83. Turro, N. J. *Modern Molecular Photochemistry*. Benjamin/Cummings Publishing Company: Menlo Park, CA, 1978.

84. Hänggi, P.; Talkner, P.; Borkovec, M. Reaction-rate theory: fifty years after Kramers. *Reviews of modern physics* **1990,** *62*, 251.

85. Pauling, L.; Corey, R. B.; Branson, H. R. The Structure of Proteins - 2 Hydrogen-Bonded Helical Configurations of the Polypeptide Chain. *P Natl Acad Sci USA* **1951,** *37*, 205-211.

86. Alberts, B.; Johnson, A.; Lewis, J.; Raff, M.; Roberts, K.; Walter, P. The shape and structure of proteins. **2002**.

87. Finn, R. D.; Mistry, J.; Schuster-Bockler, B.; Griffiths-Jones, S.; Hollich, V.; Lassmann, T.; Moxon, S.; Marshall, M.; Khanna, A.; Durbin, R., et al. Pfam: clans, web tools and services. *Nucleic Acids Res* **2006,** *34*, D247-D251.

88. Sonnhammer, E. L. L.; Eddy, S. R.; Birney, E.; Bateman, A.; Durbin, R. Pfam: multiple sequence alignments and HMM-profiles of protein domains. *Nucleic Acids Res* **1998,** *26*, 320-322.

89. Bateman, A.; Martin, M. J.; O'Donovan, C.; Magrane, M.; Alpi, E.; Antunes, R.; Bely, B.; Bingley, M.; Bonilla, C.; Britto, R., et al. UniProt: the universal protein knowledgebase. *Nucleic Acids Res* **2017,** *45*, D158-D169.

90. Wu, C. H.; Yeh, L. S. L.; Huang, H. Z.; Arminski, L.; Castro-Alvear, J.; Chen, Y. X.; Hu, Z. Z.; Kourtesis, P.; Ledley, R. S.; Suzek, B. E., et al. The Protein Information Resource. *Nucleic Acids Res* **2003,** *31*, 345-347.

91. Boeckmann, B.; Bairoch, A.; Apweiler, R.; Blatter, M. C.; Estreicher, A.; Gasteiger, E.; Martin, M. J.; Michoud, K.; O'Donovan, C.; Phan, I., et al. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res* **2003,** *31*, 365-370.

92. Bairoch, A.; Apweiler, R. The SWISS-PROT protein sequence data bank and its new supplement TREMBL. *Nucleic Acids Res* **1996,** *24*, 21-25.

93. Bairoch, A. Serendipity in bioinformatics, the tribulations of a Swiss bioinformatician through exciting times! *Bioinformatics* **2000,** *16*, 48-64.

94. O'Donovan, C.; Martin, M. J.; Gattiker, A.; Gasteiger, E.; Bairoch, A.; Apweiler, R. High-quality protein knowledge resource: SWISS-PROT and TrEMBL. *Briefings in bioinformatics* **2002,** *3*, 275-284.

95. Asgari, E.; Mofrad, M. R. K. Continuous Distributed Representation of Biological Sequences for Deep Proteomics and Genomics. *Plos One* **2015,** *10*.

96. Turian, J.; Ratinov, L.; Bengio, Y. Word representations: A simple and general method for semi-supervised learning. *Acl 2010: 48th Annual Meeting of the Association for Computational Linguistics* **2010**, 384-394.

97. Guthrie, D.; Allison, B.; Liu, W.; Guthrie, L.; Wilks, Y. In *A closer look at skip-gram modelling*, Proceedings of the 5th international Conference on Language Resources and Evaluation (LREC-2006), sn: 2006; pp 1-4.

98. Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res* **2000,** *28*, 235-242.

99. Cox, D. R. The Regression-Analysis of Binary Sequences. *J Roy Stat Soc B* **1958,** *20*, 215-242.

100. Leinweber, D. J. Stupid data miner tricks: overfitting the S&P 500. *Journal of Investing* **2007,** *16*, 15.

101. Bühlmann, P.; Van De Geer, S. *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media: 2011.

102. LeCun, Y. LeNet-5, convolutional neural networks. *URL:* http://yann*. lecun. com/exdb/lenet* **2015**, 20.

103. Chollet, F., Keras. In https://keras.io, 2015.

104. Hahnloser, R. H.; Sarpeshkar, R.; Mahowald, M. A.; Douglas, R. J.; Seung, H. S. Digital selection and analogue amplification coexist in a cortex-inspired silicon circuit. *Nature* **2000,** *405*, 947-51.

105. Lawrence, S.; Giles, C. L.; Tsoi, A. C.; Back, A. D. Face recognition: A convolutional neural-network approach. *Ieee T Neural Networ* **1997,** *8*, 98-113.

106. Goodfellow, I.; Bengio, Y.; Courville, A.; Bengio, Y. *Deep learning*. MIT press Cambridge: 2016; Vol. 1.

107. Werbos, P. J. *The roots of backpropagation: from ordered derivatives to neural networks and political forecasting*. John Wiley & Sons: 1994; Vol. 1.

108. Bengio, Y.; Simard, P.; Frasconi, P. Learning Long-Term Dependencies with Gradient Descent Is Difficult. *Ieee T Neural Networ* **1994,** *5*, 157-166.

109. Keeler, J. D.; Hartman, E. J.; Liano, K.; Ferguson, R. B., Residual activation neural network. Google Patents: 1996.

110. Dunker, A. K.; Lawson, J. D.; Brown, C. J.; Williams, R. M.; Romero, P.; Oh, J. S.; Oldfield, C. J.; Campen, A. M.; Ratliff, C. R.; Hipps, K. W., et al. Intrinsically disordered protein. *J Mol Graph Model* **2001,** *19*, 26-59.

111. Dyson, H. J.; Wright, P. E. Intrinsically unstructured proteins and their functions. *Nat Rev Mol Cell Bio* **2005,** *6*, 197-208.

112. Dunker, A. K.; Silman, I.; Uversky, V. N.; Sussman, J. L. Function and structure of inherently disordered proteins. *Curr Opin Struc Biol* **2008,** *18*, 756-764.

113. Romero, P.; Obradovic, Z.; Li, X.; Garner, E. C.; Brown, C. J.; Dunker, A. K. Sequence complexity of disordered protein. *Proteins: Structure, Function, and Bioinformatics* **2001,** *42*, 38-48.

114. https://mcule.com/apps/1-click-docking/ 1-Click Docking.

115. Irwin, J. J.; Shoichet, B. K.; Mysinger, M. M.; Huang, N.; Colizzi, F.; Wassam, P.; Cao, Y. Q. Automated Docking Screens: A Feasibility Study. *J Med Chem* **2009,** *52*, 5712-5720.

116. Kuntz, I. D.; Blaney, J. M.; Oatley, S. J.; Langridge, R.; Ferrin, T. E. A Geometric Approach to Macromolecule-Ligand Interactions. *J Mol Biol* **1982,** *161*, 269-288.

117. Ambros, V. The functions of animal microRNAs. *Nature* **2004,** *431*, 350-355.

118. Bartel, D. P. MicroRNAs: Genomics, biogenesis, mechanism, and function. *Cell* **2004,** *116*, 281-297.

119. Lewis, B. P.; Shih, I. H.; Jones-Rhoades, M. W.; Bartel, D. P.; Burge, C. B. Prediction of mammalian microRNA targets. *Cell* **2003,** *115*, 787-798.

120. Lewis, B. P.; Burge, C. B.; Bartel, D. P. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell* **2005,** *120*, 15-20.

121. Ellwanger, D. C.; Buttner, F. A.; Mewes, H. W.; Stumpflen, V. The sufficient minimal set of miRNA seed types. *Bioinformatics* **2011,** *27*, 1346-1350.

122. Wu, T. P.; Ruan, K. C.; Liu, W. Y. A fluorescence-labeling method for sequencing small RNA on polyacrylamide gel. *Nucleic Acids Res* **1996,** *24*, 3472-3473.

123. Chan, H. M.; Chan, L. S.; Wong, R. N. S.; Li, H. W. Direct Quantification of Single-Molecules of MicroRNA by Total Internal Reflection Fluorescence Microscopy. *Anal Chem* **2010,** *82*, 6911-6918.

124. Campos-Melo, D.; Droppelmann, C. A.; Volkening, K.; Strong, M. J. Comprehensive Luciferase-Based Reporter Gene Assay Reveals Previously Masked Up-Regulatory Effects of miRNAs. *Int J Mol Sci* **2014,** *15*, 15592-15602.

125. Jin, Y.; Chen, Z.; Liu, X.; Zhou, X. Evaluating the microRNA targeting sites by luciferase reporter gene assay. In *MicroRNA Protocols*, Springer: 2013; pp 117-127.

126. Clement, T.; Salone, V.; Rederstorff, M. Dual Luciferase Gene Reporter Assays to Study miRNA Function. *Methods Mol Biol* **2015,** *1296*, 187-198.