We want to investigate how the presence of water affect the correlator computed from x-ray scattering data. We believe that correlating scattering events from pairs of water molecules may be responsible for the u-shape background we see in the DNA data. Here I present an attempt to compute such correlator by simulating water with molecular dynamics (MD) simulations.

1 Direct calculation of the correlator from simulation

The intensity from scattering from a pair of water molecules is:

$$I(\vec{q},\omega) = \left| \sum_{i}^{2} f(q) e^{-i\vec{q}\cdot(R_\omega \vec{r_i})} \right|^2 \tag{1}$$

where R_ω is the rotation operator by angle ω and $f(q)$ is the atomic form factor for oxygen. I am ignoring hydrogen for now since it does not scatter x-ray strongly. If we just expand equation 1 explicitly, we have

$$I(\vec{q},\omega) = 2\big|f(q)\big|^2 (1 + \cos \vec{q} \cdot (R_\omega \vec{r_{12}})). \tag{2}$$

We want to compute, from the simulation data, the correlator between two pairs of water molecule, i.e. a single 4-point ensemble. The correlator is the following integral:

$$C(\vec{q_1},\vec{q_2}) = \int_\omega I(\vec{q_1},\omega) I(\vec{q_2},\omega) d\omega. \tag{3}$$

Assuming our water box simulation is an good approximation of the statistical behavior of real water, we can estimate $C(\vec{q_1},\vec{q_2})$ by summing over all possible 4-point ensemble (tetrahedrons) in the simulations. Specifically,

$$C(\vec{q_1},\vec{q_2}) = \frac{4\big|f(q)\big|^4}{N_{tthd}} \sum_{m=1}^{M} \sum_{i,j,k,l}^{N} (1 + \cos \vec{q_1} \cdot \vec{r_{ij,m}})(1 + \cos \vec{q_2} \cdot \vec{r_{kl,m}}) \tag{4}$$

where N_{tthd} id the total number of tetrahedrons in M simulations frames, N is the total number of water molecules in the simulation, the index m denotes the m-th frame in the simulation. We are averaging over M statistically independent frames. The summation over indices i, j, k, and l are over all unique 4-point ensembles.

The dependence on ω is removed going from equation 3 to 4 as the MD simulation takes care of sampling over all possible orientation of tetrahedrons. The assumption here is that the water model used in the MD simulation is a reasonable and we have enough statistically independent frames from the MD trajectory.

The atomic form factor for oxygen is approximately the sum of some gaussian functions. With $0 < q < 25\text{Å}^{-1}$, $f(q)$ is

$$f(q) = \sum_{i}^{4} a_i \exp\left(-b_i(\frac{q}{4\pi})^2\right) + c \tag{5}$$

and the a_i's and b_i's are constants summarized in the table below.

a_1	b_1	a_2	b_2	a_3	b_3	a_4	b_4	c
3.0485	13.2771	2.2868	5.7011	1.5463	0.3239	0.867	32.9089	0.2508

2 Method of computation

The computation is done in three steps: 1) MD simulation; 2) finding unique tetrahedrons formed by four water molecules; 3) computing C according to equation 4.

2.1 MD simulation

We use GROMACS version 5.0.2 to run the MD simulations. We start with a single water molecule, define a cubic simulation box that is 2.2 nm on each side, and then use the build-in solvation function in GROMACS to fill the box with water molecules. The automatic solvation results in 339 water molecules in the box. We use an AMBER force field (amber99sb-ildn) and a TIP3P model for water molecules. Periodic boundary condition is also applied. Specific simulation parameters are saved in nvt-pr-md_RUNNAME.mdp files in /home/shenglan/MD_simulations/water_box on zauber. Every 1 ps in simulation time, the coordinates of the water molecules are recorded. The trajectories ranges from 100 frames to 10000 frames (100 ps to 10 nm simulation times).

2.2 Defining tetrahedrons

From the MD simulations, we can build a set water tetrahedrons. The position of a water molecule is defined only by the oxygen atom since the hydrogen atoms scatter x-ray much more weakly. In every simulation frame, we can find the three nearest neighbors for every water molecules and form one tetrahedron this way. Using only the nearest neighbors, which probably contribute most to C than neighbors further away, we can reduce the the number of tetrahedrons we need to sample and thus computation time. There are approximately as many unique[1] nearest-neighbor tetrahedrons as water molecules per simulation frame. For instance, 1000 statistically independent simulation frames generate about 330k tetrahedrons.

[1]The combination of four water molecules that make up the vertices of the tetrahedron is unique. Some simulation frames yield fewer than 339 tetrahedrons because two molecules that are nearest neighbors of each other happen to share the two other nearest neighbors.

If the simulation is true the actual physics of liquid water, this set of water tetrahedron is a representation of the true distribution of tetrahedron geometries that are possible for water. Its underlying statistics are therefore a good estimate of those of the true distribution. One way to check is to look at the radial distribution function, i.e. the distribution of distances between pairs of water molecules (oxygens); it has a well-known shape. Figure 1 shows the radial function produced by looking at as simulation of 10000 frames and 339 water molecules in $2.2 \times 2.2 \times 2.2$ nm box. The shape looks reasonable with a peak at about 0.3 nm. It will be interesting to compare radial distribution function produced by systems of different sizes since my understanding is that the periodic boundary condition has a larger effect on smaller systems.

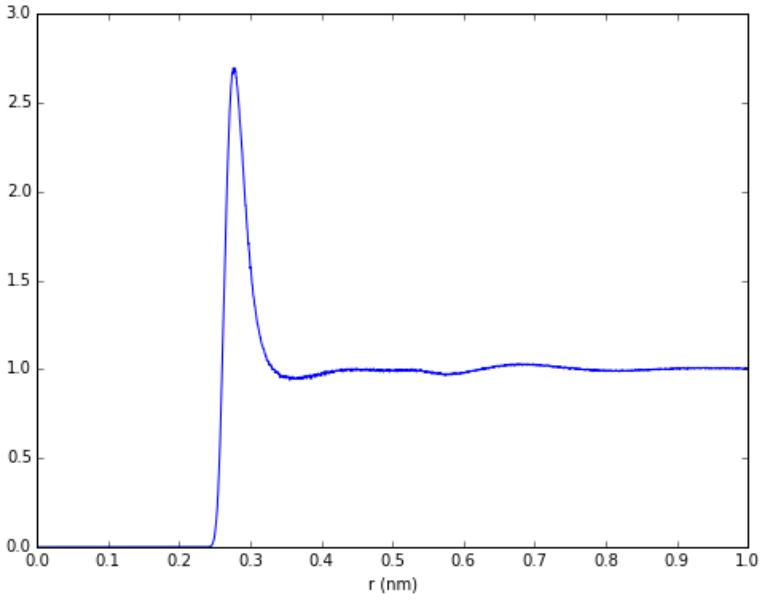

Figure 1: Radial distribution function of simulated water.

Look into the some statistics of the tetrahedrons (e.g. length of sides) and write about them here.

2.3 Computing the correlator

The sampled tetrahedrons coordinates provide r_{ij} and r_{kl} in equation 4. We then need to choose $\vec{q_1}$ and $\vec{q_2}$. Specifically, we can fix $\vec{q_1}$ and then define $\vec{q_2}(|q_2|, \phi)$, where ϕ is the angle between the two vectors on the detector plane. For autocorrelation, the magnitude of both vectors is $|q|$. For $\vec{q} = (q_x, q_y, q_z)$,

$$q_x = |q| \sqrt{1 - (\frac{|q|}{2q_{beam}})^2} \cos(\phi) \tag{6}$$

$$q_y = |q| \sqrt{1 - (\frac{|q|}{2q_{beam}})^2} \sin(\phi) \tag{7}$$

$$q_z = -\frac{|q|^2}{2q_{beam}} \tag{8}$$

In the case of the autocorrelator, we will fix ϕ for $\vec{q_1}$ at zero and generate $\vec{q_2}(\phi)$ ($\phi \in [0, \pi]$) . In the case of different magnitudes of $\vec{q_1}$ and $\vec{q_2}$, $|q|$ in the equations above are replaced by $|q_1|$ and $|q_2|$.

I personally find it more intuitive to think about the inverse of $|q|$, i.e. the length scale we are probing with the x-ray scattering. As a test, I computed the autocorrelators for the following lengths in nm: 0.23, 0.292, 0.306, 0.322, 0.34, 0.36, 0.38, 0.405, 0.433, 0.465. In terms of $\mathring{A}^{-1}$, the magnitude of the q-vectors are 2.73, 2.15, 2.05, 1.95, 1.85, 1.75, 1.65, 1.55, 1.45, 1.35.

3 Results and open questions

Figures 2 and 3 summarizes the results for computing autocorrelations for the range of $|q|$ listed in the previous section. Figures 2 shows autocorrelators as a function of $\cos\psi$, where ψ is the angle between $\vec{q_1}$ and $\vec{q_2}$. The minimums of the curves are aligned to show their shapes better. At smaller $|q|$, the autocorrelator is clearly u-shape and symmetric about $\cos\psi = 0$. At larger $|q|$, two peaks appear and some asymmetry becomes more prominent. The peaks can be intuitively understood as more atomic details at shorter length scales as we examine the autocorrelator increasing $|q|$. I attribute the asymmetry, which should not appear if the tetrahedrons are properly construct, to some kind of under or biased sampling of the water molecule configurations.

Figures 3 shows the relative magnitude of the autocorrelator at different $|q|$.The autocorrlators are lrger in magnitude when $|q|$ is smaller.

Some aspects of the computation are still not well-understood by me. The sampling of tetrahedron is done by modeling the water with MD simulations. Does this sampling truly represent configurations of water molecules? One indirect test is to see if the autocorrelator as a function of $\cos\psi$ for a given $|q|$ converges as more water molecule configurations are sampled by the simulations. Figure 4 tries to visualize the converge: for each $|q|$, the autocorrelator is computed with increasing number of simulations frames until data from all 10000 frames are used. The difference between consecutive curves are then calculated for each $|q|$ and plotted as a function of number of simulation frames used. This plot suggests that as more water molecule configurations are sampled, the differences between consecutive curves tends to zero, i.e. the autoccorrelator tends to stabilize to a particular functional form.

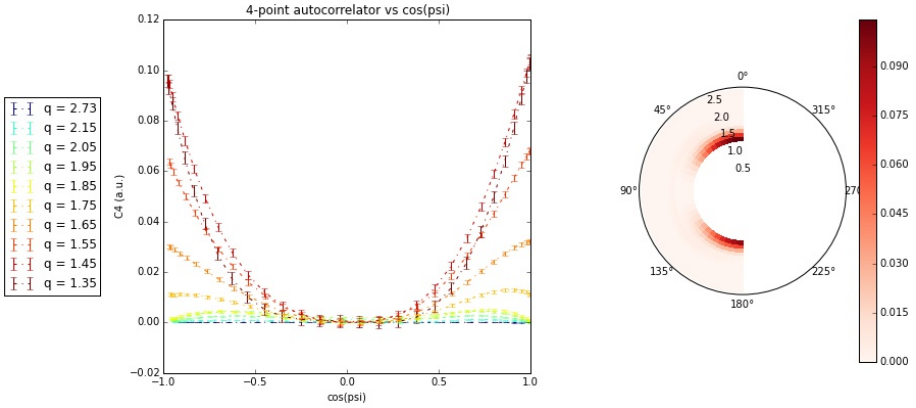

Figure 2: Autocorrelator for various magnitudes for $\vec{q}$. The curves in the plot on the left are aligned by their minimums in order to show the change in their shape as a function of q. The plot on the right is the polar version of the same information.

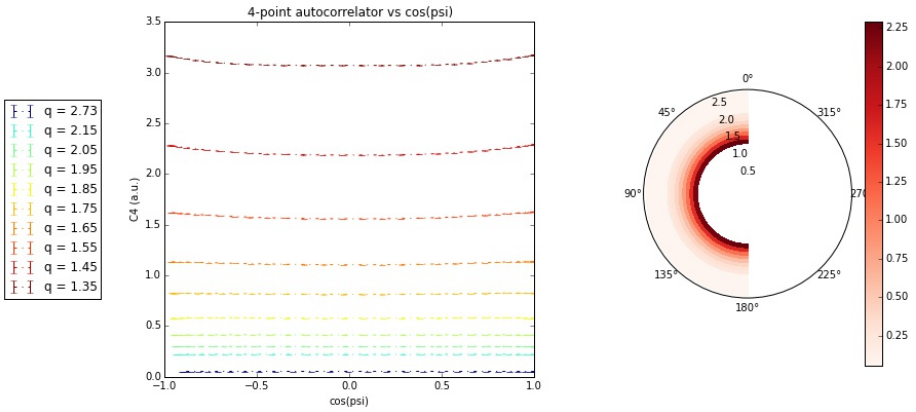

Figure 3: Similarly to figure 2, autocorrelator for various magnitudes for $\vec{q}$. The minimums of the curves are not aligned by their minimums in order to show the relative magnitude of the autocorrelator as a function of q.

My suspicion is that this convergence is not achieved at the same rate for different $|q|$'s; at large $|q|$'s, the autocorrelator looks less symmetric as compared to that at smaller $|q|$'s (figure 5). My intuition is that this can probably be explained by the the radial distribution function (figure 1) drops off rapidly to zero at about $r < 0.3$ nm. Probing at length scales than 0.3 nm ($|q| > 2.1\mathring{A}^{-1}$) will require more simulation frames to reach convergence.

There are more explorations that might be worthwhile to explore. First, I should establish a better understanding of the statistics of the geometric prop-

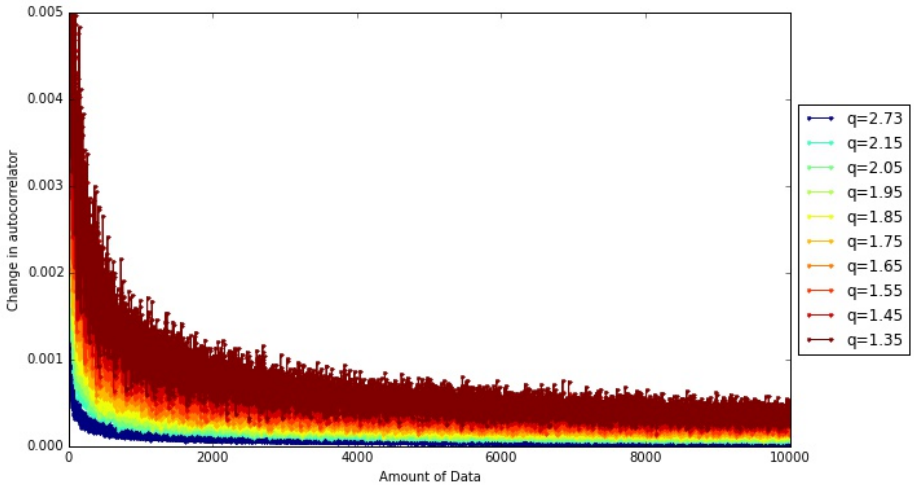

Figure 4: The differences between consecutive autocorrelators tend to zero as more simulation data are used to compute them.

erties of the 3 million water tetrahedrons I have obtained from MD simulations so far; I should also try to figure out what the statistics should be for real water. Also, I want to further explore the effect of simulation box size on the distribution of tetrahedrons and on the computed correlator. I want to try Derek's code (thor) and compare its results to correlators computed from MD simulations. Moving on from here, I should also explore how this simulation can help understanding the DNA data.

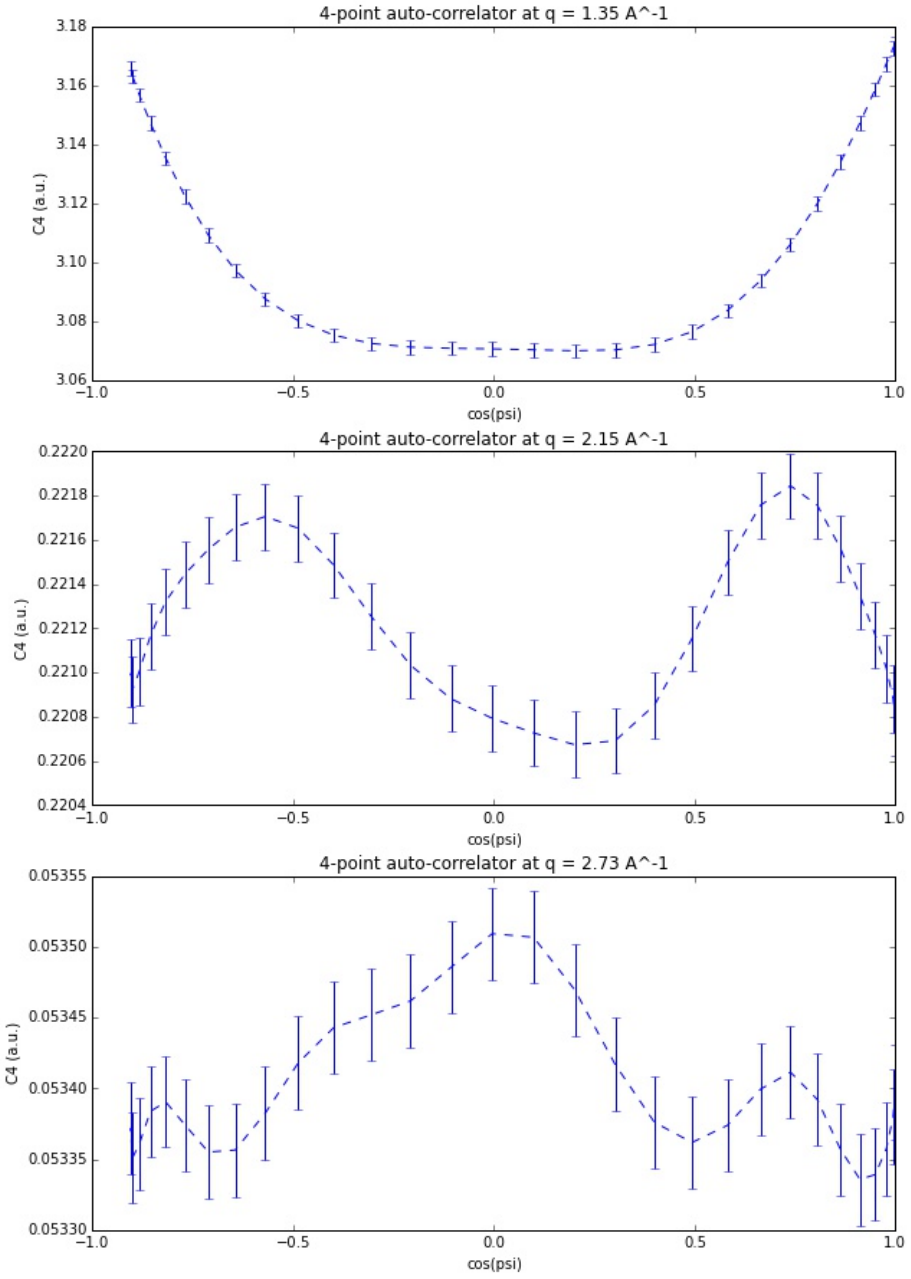

Figure 5: Comparing the shapes of the autocorrelator curves at different $|q|$'s