INVESTIGATING THE VOLUME PERTURBATION OF NEURAL LIPID-MEMBRANE BY PROTEIN IN THE INITIAL STAGES OF ALZHEIMER'S DISEASE

SANTONA TULI WITH PROJECT COLLABORATORS GARNER COCHRAN AND AUSTIN WORKMAN

Date: December 11, 2012.

1. Introduction

In the Fall of 2012 we worked on a project which involved analysing data from molecular dynamics simulations produced in Dr. Kelvin Cheng's (Trinity University Department of Physics) research on understanding the initial stages of Alzheimer's disease. The primary articles which inspired our work are *Analysis of Lipid Surface Area in Protein-Membrane Systems Combining Voronoi Tessellation and Monte Carlo Integration Methods* by Takahura Mori, Fumiko Ogushi, and Yuji Sugita, *Molecular Dynamics Simulations Reveal the Protective Role of Cholesterol in β-Amyloid Protein-Induced Membrane Disruptions in Neuronal Membrane Mimics* by Liming Qiu, Creighton Buie, Andrew Reay, Mark W. Vaughn and Kwan hong Cheng, and *GridMAT-MD: A Grid-Based Membrane Analysis Tool for Use With Molecular Dynamics* by William J. Allen, Justin A. Lemkul and David R. Bevan.

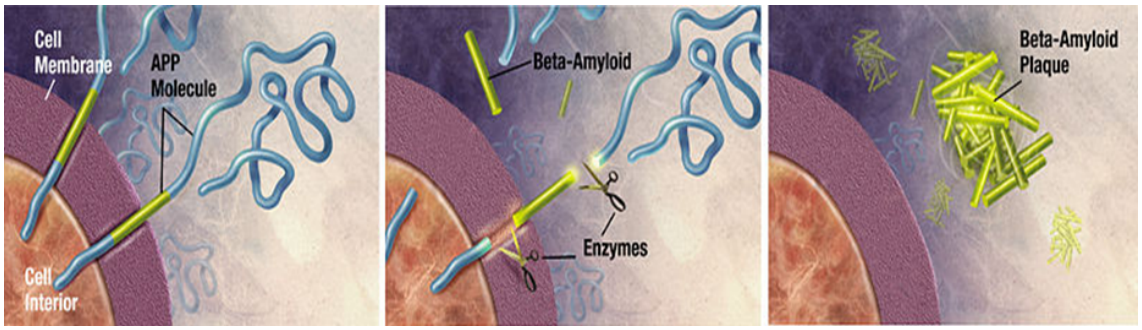
We learned from a comprehensive literature search and from various discussion sessions with Dr. Cheng and Dr. Nguyen that the technique of molecular dynamics simulation has become a powerful research tool to investigate structural properties of biological and chemical systems as well as interactions between biological membranes on a mesoscopic level. Often in biology, we need to resort to simulations as the only approach to understanding certain systems, since experimentation is difficult or impossible due to the time scales involved or lack of predictability. For instance, since one cannot know in advance when a disease such as Alzheimer's may set in, it would be impossible to investigate the origins of the disease by examining a patient. In fact, we hope that some day, thanks to molecular dynamics simulations and their analysis, we may be able to in fact predict the likelihood of a disease in an individual so that we may take adequate preventive measures. Molecular dynamics is an awe-inspiring tool that enables researchers to begin to explore interactions invisible to the human eye due to size of particles and extremely short time-scales.

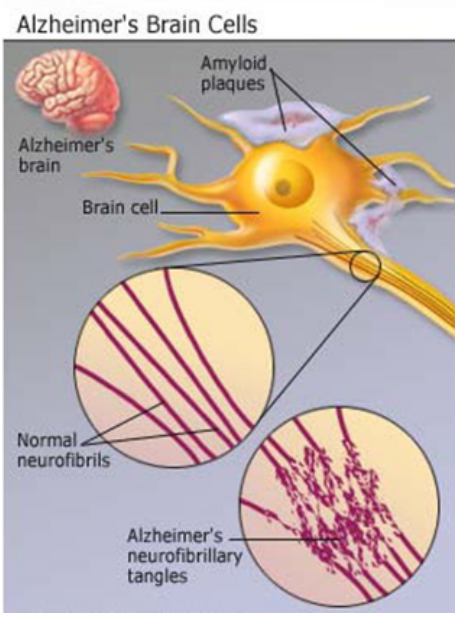
The overall goal of the project is to understand the process through which APP (a trans-membrane protein) is divided into fragments including abnormally folded beta-amyloid which accumulate to form senile plaques, thereby initiating the onset of Alzheimer's disease (please refer to subsection Alzheimer's Disease). My involvement in this project has been to perform extensive literature search on the disease as well as on the simulation methods, understand and run a program called Surround written in C in order to determine the Annular region of a protein molecule, generate the appropriate volume perturbation graphs (for lipids, cholesterol and water) for data analysis, and prepare and lead the project progress and final presentations.

The paper by Takahura Mori et al. has the most advanced method of combining the methods of Monte Carlo integration and Voronoi tessellations, in order to simulate protein-lipid interactions at a membrane. These methods are explained below. The paper by Liming Qiu et al. illustrates the methods of analysis of results from a simulation similar to ours, and the paper by William J. Allen et al. illustrates a different simulation technique.

What follows is a discussion of the project and our progress through the semester as a whole, with particular attention to the discussion of my contribution.

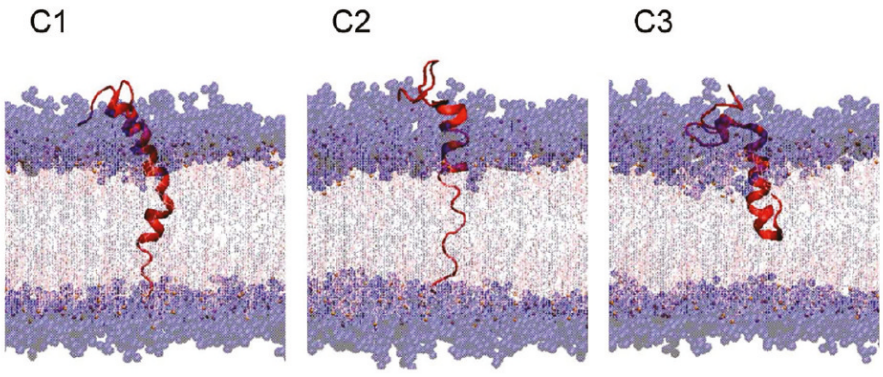
1.1. **Alzheimer's Disease.** Alzheimer's disease is a form of dementia which typically affects people 65 years of age or older. Unfortunately, currently there does not exist any cure for the disease; it worsens as it progresses eventually leading to death. As a result, there is a significant amount of ongoing research as to the origins of the disease. If we can understand the disease better we may be able to come up with preventive measures. Alzheimer's Disease is a protein misfolding disease. The accumulation of abnormally folded beta- and tau-amyloid proteins outside the neural membrane in our brain causes the symptoms of Alzheimer's disease. Neurons are surrounded by a plasma membrane consisting of a bilayer of lipid molecules with protein structures embedded in it. Normatively, transmembrane proteins penetrate though the neuron membrane to perform their necessary functions. Amyloid precursor protein (APP) is a transmembrane protein critical to neuron growth and repair. In certain cases (those that lead to Alzheimer's disease) an unknown process causes the APP to divide into smaller portions such as tau- and beta-amyloids. Fibrils of beta-amyloids clump together outside neurons and form senile plaques. These senile plaques have no beneficial purpose, and eventually lead to the symptoms which are then recognized as Alzheimer's disease. The two figures below help to illustrate this process.





The goal of the research in this field is therefore to understand the process through which APP is divided into fragments including beta-amyloid. This process is called misfolding. We achieve this goal through simulating protein-lipid interactions using molecular dynamics. From the simulation results we can produce volume perturbation graphs for the different kinds of molecules. The procedures we use are outlined below. Our final graphs are included in the results section at the end of this report, however the continuation of our work next semester will help us improve on these results and adequately interpret them.

1.2. **Our Modelling Approach.** In particular, we analyse the data from the simulation of the instance called C2 shown in the figure below (C1 and C3 are also shown for comparison).



We perform two different kinds of initial analysis. First, we use surround to take the Protein Data Bank (PDB) file for C2 and use it to generate an annular region for the protein molecules and determine which lipid molecules belong to this region. Next we use a combination of Monte-Carlo Integration and Voronoi Tessellation (in the program VORO++) to determine which lipid, cholesterol and water molecules in

C2 belong in close proximity to the protein molecules. Since the two programs use different procedures to work, we expect slightly different results. Indeed, we find that the outputs are somewhat different. However, there is overlap between the outputs and this assures us we are on the right track. We then pick the method we deem more accurate, which in this case is voronoi tessellation using VORO++ and upon further improving on the process we are finally able to generate the adequate graphs for analysis. We use volume perturbation as the technique for analysis of the simulation results.

The following sections discuss each method we employ in our modelling approach separately.

1.3. **Monte Carlo Integration.** Given a function which is difficult to integrate analytically, its the area or volume (integral) can be transformed into an easier computation which relies on random point generation:

Let $X = \{x_1, x_2, \ldots, x_N\}$ be randomly generated points across a sample size. The volume integral becomes:

$$\int f \mathrm{d}V \approx V\langle f \rangle \pm V\sqrt{\frac{\langle f^2 \rangle - \langle f \rangle^2}{N}}$$

Where $\langle f \rangle = \frac{1}{N}\sum_{i=1}^{N} f(x_i)$ and $\langle f^2 \rangle = \frac{1}{N}\sum_{i=1}^{N} f^2(x_i)$

This helps us with the process of voronoi tessellation which is explained below.

1.4. **Voronoi Tessellation.** Voronoi Tessellation is defined as The partitioning of a plane with n points into convex polygons such that each polygon contains exactly one generating point and every point in a given polygon is closer to its generating point than to any other.

Using these methods in relation to protein lipid interaction is to simply put N random points over a large area of lipids and proteins and try to make a structure of one proteins interaction with one lipid specifically. The voronoi tessellation performed by VORO++ (see below) uses the monte carlo method to find area and subsequently depth and volume of the molecules we provide it with. It uses the voronoi method to produce a structure of each kind of molecules.

Given the locations of the atoms in the C2 protein-lipid interaction, we model the area that these atoms encompass using the 3-d voronoi tessellation method. The 3-d voronoi tessellation method works by finding the "neighborhood" of every atom. A more accurate description is given a point $P_i \in \mathbb{R}^3$, the region associated with such a point, call this region R_i is defined as:

$$R_i = \left\{ c \in \mathbb{R}^3 | d(c, P_i) \leq d(x, P_j) \forall j \neq i \right\}$$

Note that in this definition, the distance function d is the euclidean distance.

We then take these regions and consider the regions encompassed by multiple voronoi regions associated by either the kind of atom they represent or the kind of molecule they represent. Once we are able to categorize these areas, we perform statistical analysis on their regions by trying to model how much the volumes in certain regions were changed by the insertion of the protein. This is known as volume perturbation.

2. Surround

We first use the program surround (in programming language C) written by Dr. Kelvin Cheng and group at Texas Tech University, to separate a pdb file of our C2 configuration (a system containing protein, lipids, and solvent molecules) into an annular lipid (AL) region and a non-annual lipid (nAL) region. The program takes the pdb file as its input and other parameters may be specified. For instance, R is the parameter for controlling the size of the region over which the annular lipids are searched for and defined. That is, for any given value of R, surround searches for lipid atoms in the R neighborhood of every protein atom in the system. We set the default value of $R = 10$ Å, and we are able change its value depending on our desired precision of protein-lipid interface and our desired region. The parameter δ controls the spatial resolution of the system and its default value is set to 1 Å. As surround searches the R neighborhood of every protein atom for lipid atoms, it matches the detected lipid atoms to lipid molecules (indexed in the pdb file). Once it finds more than 50% of any given lipid molecules atoms in the R neighborhood of a protein atom, it considers that lipid molecule as a whole to be within the interaction range of the protein atom. That is, it considers that lipid molecule to be an annular lipid for that protein atom. It repeats this process with every protein atom in the system and finally characterizes and lists all the lipid molecules that fall within the overall annular zone.

Finally, the parameter include is used to command surround to create either the AL region or the nAL region. If include takes its default value, which is 0, the program will create the AL region; otherwise, it will create the nAL region. Parameter PCcount sets the selection criterion for POP (lipid) molecules in the system. The default value of this parameter is 26, which is half of the number of atoms of a POP molecule. Therefore, the program decides that a POP molecule is in the AL region if no less than 26 of its constituent atoms are in the AL region of at least one protein atom. Changing this value to 52, for instance, causes surround to identify those POP molecules with all their atoms in the AL region (R neighborhood of a given protein atom) to belong to the overall AL region. A value of 0 will take all POP molecules in the AL region, and a value greater than 52 will make it impossible for any POP

molecules to be considered in the AL region. Similarly, the parameters CHLcount and SOLcount can be used for the same purpose for CHO (cholesterol) and SOL (solvent) molecules, respectively. The default values of CHLcount and SOLcount are 15 and 1 respectively. However, we only run surround for the lipids, since by comparison with the VORO++ results we find that VORO++ is more adequate for our further work and therefore continue with VORO++.

There are two notable disadvantages to categorizing lipids as annular or non-annular using this method employed by surround. First, the region over which the program searches for annular lipids is fairly arbitrary, as we choose the value for R. We must choose a large enough R so that all the lipid molecules that interact with our protein molecules do fall within this region we define as annular. At the same time, as we choose larger values of R we lose precision of knowing which lipid is closest to, and therefore interacts with, which protein. Secondly, since the percentage of atoms per lipid required to designate a lipid molecule as annular is cut off at 50% (a value chosen arbitrarily by the user), we may gain unnecessary lipids per protein atom, or conversely, lose interacting lipids for any given protein atom and its R neighborhood. We use a reasonably large R ($R = 10$ Å) when we run surround for lipids in the annular region.

We then employ the Voronoi tessellation method, using VORO++, outlined below to find the first generation neighboring lipid molecules for each protein molecules that is, to identify the lipids that actually share an interface with the protein molecules. In VORO++, this becomes our criterion for annular lipids. Therefore, surround acts as a first approximation for our annular region, which we then compare with the VORO++ annular region to determine which is more comprehensive.

3. Understanding the PDB Format

A typical PDB format file will contain atomic coordinates for a diverse collection of proteins, small molecules, ions and water. Each atom is entered as a line of information that starts with a keyword: either ATOM or HETATM. By tradition, the ATOM keyword is used to identify proteins or nucleic acid atoms, and keyword HETATM is used to identify atoms in small molecules. Following this keyword, there is a list of information about the atom, including its name, its number in the file, the name and number of the residue it belongs to, one letter to specify the chain (in oligomeric proteins), its x, y, and z coordinates, and an occupancy and temperature factor (desc ribed in more detail below).

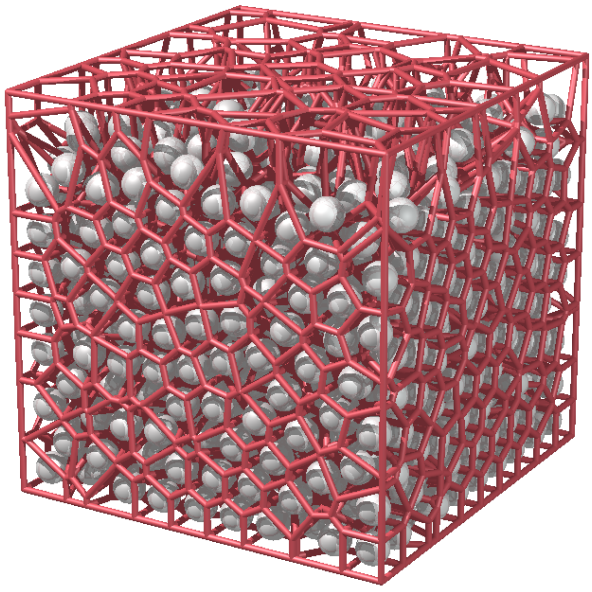
The information given provides a lot of control when exploring any structure. For instance, most molecular graphics programs enable us to color identified portions of the molecule selectively- It allows us in our project to only view a certain interaction between lipid and protein by taking out water and other unwanted parts of the model.

Below is an example of a couple of atoms indexed in PDB format.

```
ATOM      1  N    ASP      1       76.610  77.690  57.950  1.00  0.00
ATOM      2  H1   ASP      1       77.480  77.250  57.910  1.00  0.00
ATOM      3  H2   ASP      1       76.650  78.560  57.420  1.00  0.00
ATOM      4  H3   ASP      1       76.390  77.910  58.910  1.00  0.00
ATOM      5  CA   ASP      1       75.760  76.680  57.360  1.00  0.00
ATOM      6  CB   ASP      1       76.560  76.100  56.210  1.00  0.00
ATOM      7  CG   ASP      1       75.680  74.960  55.590  1.00  0.00
ATOM      8  OD1  ASP      1       76.250  73.900  55.390  1.00  0.00
ATOM      9  OD2  ASP      1       74.480  75.230  55.390  1.00  0.00
ATOM     10  C    ASP      1       75.450  75.600  58.410  1.00  0.00
ATOM     11  O    ASP      1       76.380  75.100  59.020  1.00  0.00
ATOM     12  N    ALA      2       74.140  75.300  58.630  1.00  0.00
ATOM     13  H    ALA      2       73.420  75.920  58.250  1.00  0.00
ATOM     14  CA   ALA      2       73.740  74.130  59.470  1.00  0.00
ATOM     15  CB   ALA      2       72.220  74.040  59.410  1.00  0.00
ATOM     16  C    ALA      2       74.350  72.830  59.010  1.00  0.00
ATOM     17  O    ALA      2       74.890  72.110  59.830  1.00  0.00
```

4. Voro++

Voro++ is a program in C++ which is able to map out 3d voronoi tessellations and give the volume of the cells which encompass each atom. Voro++ also has many useful outputs, like the neighbors of each cell and the definition of the walls of each cell. For simpler examples the voro++ program will output a visualization of the data, an example of which is below.



We take the PDB file for C2 and extract the x,y, and z coordinates of the atoms in the protein-lipid interaction. In extracting the data we use an indexed list of the atoms in the PDB file which is indexed $1, \ldots, n$ so that we can cross reference it.

From this indexed list we extract another indexed list including just the index number, and the x, y, and z coordinates. We then take these coordinates and run the voro++ program on them. This outputs the volume of the voronoi cells and using the cross reference we are able to perform analysis. That is, when producing volume perturbation graphs (volume vs. distance) we are able to use the same indexing for the distance of the atoms and molecules.

5. Comparing Surround and VORO++ Outputs

The results from the two programs for the annular region are rather similar. All of the annular lipids from surround are contained in those from the Voro++ simulation. However, The Voro++ simulation contains more of these annular lipids. Since there is no mathematical deduction of comparative accuracy, we use physical understanding of the interactions and knowledge of simulation techniques and shortcomings to determine accuracy. We determine that the results from VORO++ are more accurate because they contain all the surround output molecules, and some extra ones which surround may ahve missed due to its somewhat arbitrary parameter designations. Overall, the two techniques provide cross check for each other, and we decide to proceed with Voro++.

Next, we need to perform volume perturbation analysis. For this we must define distance from the protein molecule.

6. Defining Distance

This semester we defined distance as the *average atomistic distance* in the following way:

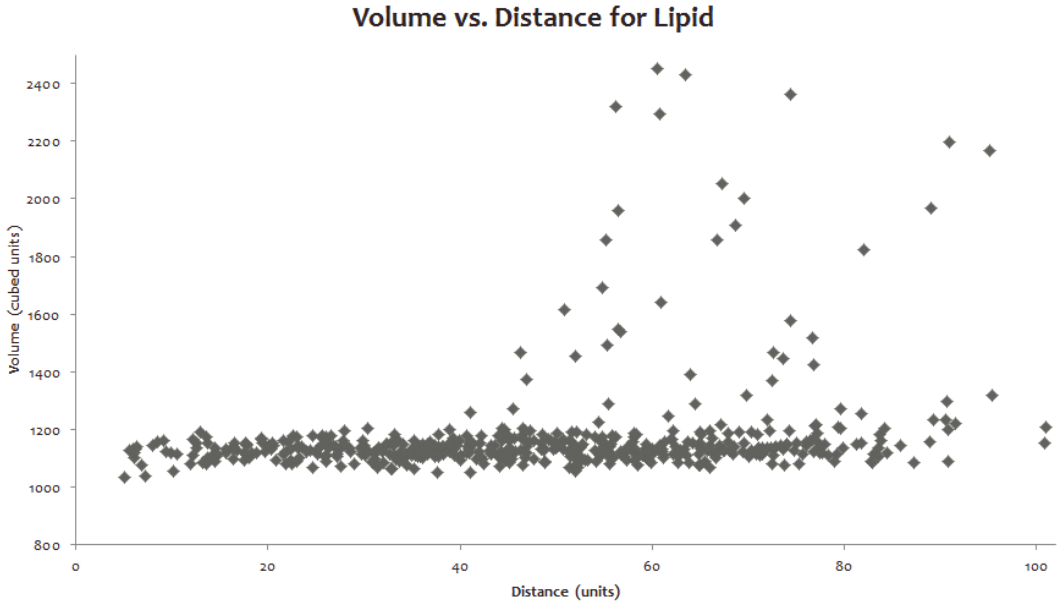
For each atom in a given (lipid, cholesterol or water) molecule, we determine the atom of the protein closest to it, and compute this distance. Next, we find the average of these distances for all the atoms in the given molecule. This is the distance between the given lipid, cholesterol or water molecule to its nearest protein molecule.

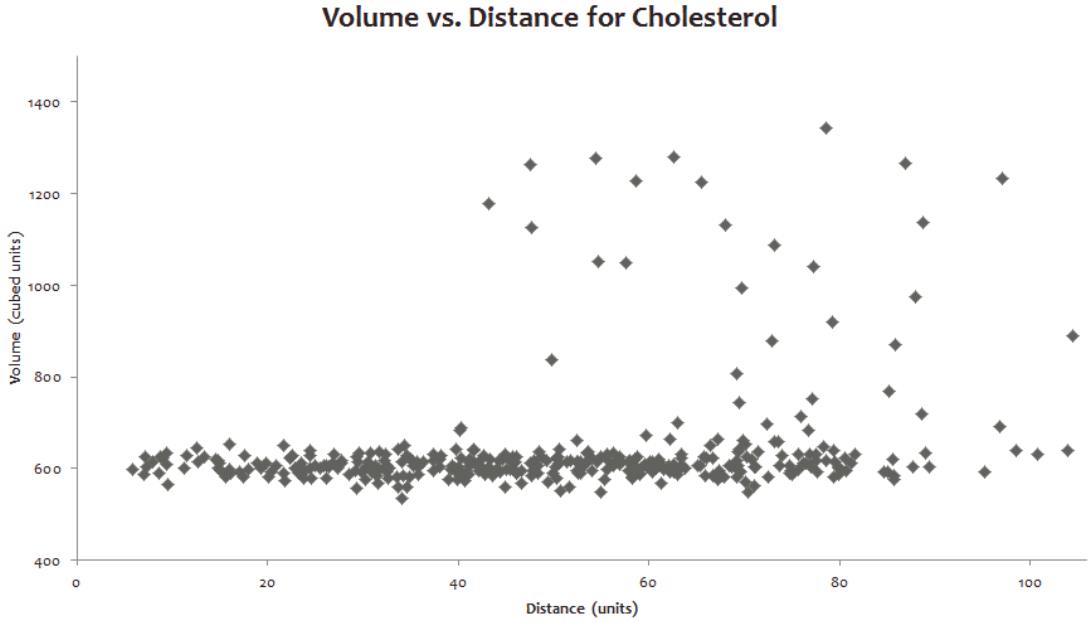
Using this definition of distance, we produce an indexed file of distances of each atom from the protein molecule and are therefore able to produce volume perturbation graphs.

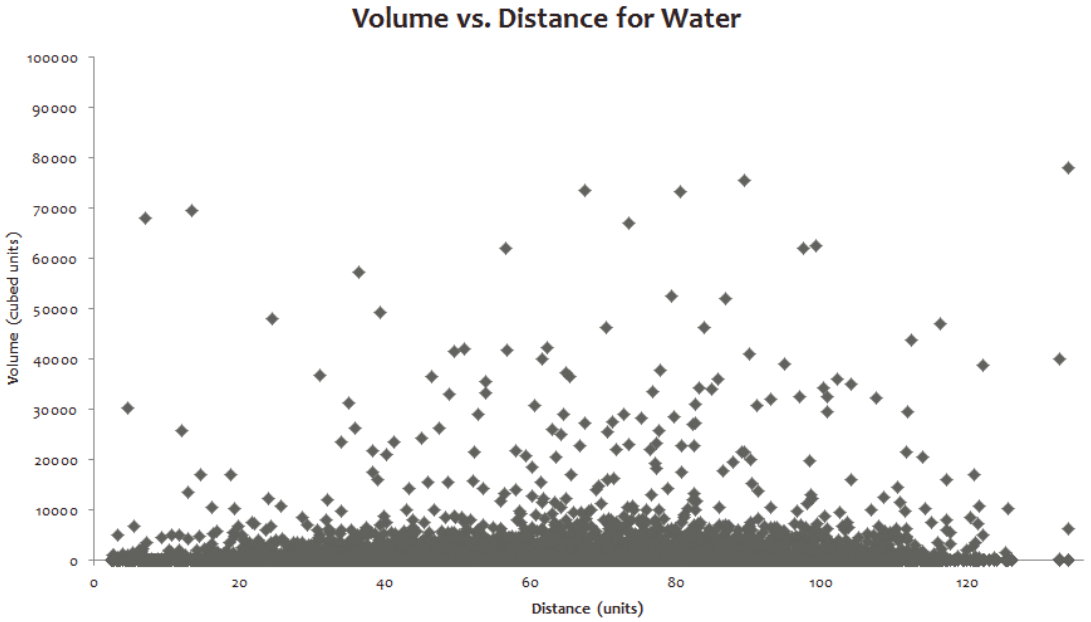
In the future, we will consider the n^{th} neighborhood distance.

7. Results and Analysis

Below are the volume perturbation plots for lipid, cholesterol and water molecules respectively:


Volume vs. Distance for Lipid
Volume (cubed units)
2400
2200
2000
1800
1600
1400
1200
1000
800
0
20
40
60
80
100
Distance (units)


Volume vs. Distance for Cholesterol
Volume (cubed units)
1400
1200
1000
800
600
400
0
20
40
60
80
100
Distance (units)



Lipids and Cholesterol: We notice that the volumes of the lipid and protein molecules are constrained for smaller distances, indicating that their volumes are perturbed for a certain range (certain radius around the protein molecule). Beyond this range the volumes seem to be fluctuating, this indicating that they are not affected uniformly by the protein which is inserting itself into the bilayer. However, we had expected the volume perturbation curve to plateaux at large distances instead of fluctuating. We understand that the fluctuations, and in particular the extremely large volumes that we obtain for certain atoms, is due to a glitch in the way we are currently using VORO++. Since we create a box which is in isolation, VORO++ assigns an arbitrarily large volume with the boundary atoms. In the future, we need to join the conjugate edges of the box so that they form a continuous (periodic) plane. Then we will produce not an isolated box but one which is a characteristic segment of an infinite but periodic volume.

Water: Since we have so many water molecules it is hard to determine a trend from these molecules. However, we are not surprised by the fact that the water molecules seem to show no volume perturbation. This is because the water molecules are much smaller compared to the lipid and protein molecules, and much more rigid in their structure and size (due to its dipole characteristics).

One interesting find of our project this semester is how fast the voronoi tessellation works. This makes us hopeful that we will be able to perform more involved analysis next semester without computational difficulties or time-constraints.

8. Future Work

Our first goal next semester, as we continue on the project, is to address all the issues that have surfaced in the result analysis from the work this semester.

Next, we plan to incorporate atomic weighting in volume generation. We also wish to consider other definitions of distance and investigate the volume perturbation differences in specifically first generation and second generation neighbors of the protein molecules.

9. Bibliography

(1) Mori, Takaharu, Fumiku Ogushi, and Yuji Sugita. *Analysis of Lipid Surface Area in Protein-Membrane Systems Combining Voronoi Tessellation and Monte Carlo Integration Methods*. Journal of Computational Chemisty (2011): n. pag. Web. 3 Dec. 2012.
(2) Goodsell, David. *Understanding PDB Data: Looking at Structures*. RCSB PDB-101. WwPDB, n.d. Web. 03 Dec. 2012.
(3) Vanden-Eijnden, Eric, and Maddalena Venturoli. *Markovian Milestoning with Voronoi Tesselations*. The Journal Chemical Physics (2009): n. pag. Jcp.aip.org. American Institute of Physics, 2009. Web. 3 Dec. 2012.
(4) Qui, Liming, Creighton Buie, Andrew Reay, Mark Vaughn, and Kwan Cheng. *Molecular Dynamics Simulations REveal the Protective Role of Cholestrol in B-Amyloid Protein-Induced Membrane Disruptions in Neuronal Membrane Mimics*. The Journal Chemical Physics (2011): n. pag. Pubs.acs.org/JPCB. Department of Physics and Engineering at Texas Tech. Web. 3 Dec. 2012.
(5) Benjamin, Bouvier, Raik Grunberg, Michael Nigles, and Frederic Cazals. *Shelling the Voronoi Interface of Protein-protein Complexes Reveals Patterns of Residue Conservation, Dynamics, and Composition*. Proteins: Structure, Function, Bioinformatics (2010): n. pag. Institut De Biologie Et De Chimie Des Proteines. Web. 3 Dec. 2012.