A3DSRINP

Advanced 3D Structural Reconstruction of Independent Nanoparticle Images Obtained from Cryogenic Transmission Electron Microscopy

Raymond Pulver
Neal Buxton
Xiaodong Wang
John Lucci
Dr. Lenore Martin
Dr. Jean-yves Herve

2016-Sp-Bioinformatics I
University of Rhode Island

Table of Contents:

1. Introduction

This section gives a general description and overview of everything included in this report. Also, the purpose and scope of this report is described, including a list of abbreviations and reference.

1. Purpose

This report is to give a detailed description of the rational and application of using an individual 2D image to reconstruct a 3D structure. It will illustrate the purpose and explain the system development. It will also explain the constraints of the program.

2. Scope

Cholesterol is carried to, from cell and transported through bloodstream by lipoproteins. There are two types of lipoproteins: lowdensity lipoprotein, or LDL, and highdensity lipoprotein, or HDL. LDL cholesterol is considered "bad" cholesterol because it can form plaque and hard deposit leading to arteries clog and make them less flexible. Heart attack or stroke will happen if the hard deposit blocks a narrowed artery. HDL cholesterol acts as a scavenger because it helps to remove LDL from the artery back to the liver.[1] Traditionally, The characteristic parameters of LDL and HDL, such as particle counts plays an important role to a better understanding and prediction of heart disease risk. But recently research suggested that geometric parameters of lipoprotein macromolecules, such as size and shape might be better correlated with cardiovascular risk than just the particle counts and density in the serum.

LDL particles are generally considered to be bad cholesterol, however, the LDL particles that fall on the large end of the LDL particle size spectrum are correlated to better health while smaller LDL particles tied to worse health. The small end of the LDL packages in the overall LDL size spectrum are considered to be an important cause of arterial plaque initiation that can eventually lead to heart attacks and stroke. And somewhat surprisingly, although HDL is conventionally considered to be "good" cholesterol, another study [2] showed that high particle counts of smaller sized HDL particles leading to a 15fold increase in the risk of heart disease. The Magnetic Resonance (NMR) spectroscopy is used as a new method to count of the number of LDL particles in presorted bin sizes but not actual sizes of individual LDL particles [3].

All of these studies conclude that more factors are at play than just the size of the particle: shape and/or other geometric distribution may be even more important, and more innovative method needs to be defined to learn the geometric distribution the LDL and HDL particles.

Cryo-Transmission Electron Microscopy (Cryo-TEM) is most frequently used for the evaluation of the ultrastructure of LDL and HDL particles. Contrasted to X-ray crystallography, which requires crystallizing the specimen, Cryo-TEM does not require a drying process. It can be used

for evaluation of additional information about the internal structure of the particles and shows in their native environment.[4] Derivation of the structure of a 3D solid object is often obtained from surface properties, applicable in this instance due to the monolayer membrane of the LDL particle and the cholesteryl ester interior. X-ray crystallography of lipoproteins is virtually nonexistent in the RCSB database. Of all PDB files analyzed, only 1000 out of 5500 residues of ApoB were able to be characterized by a homology model. The fluidity of lipoproteins and lipoproteins in general, prevents their structural analysis through traditional means. Therefore, characterization of ApoB interactions with lipid membranes requires elucidation through consecutive cycles of fitting and refitting the 3D model with cryo-TEM micrograph data. Silhouette determination of shape, as credited to Baumgart, requires estimation of the visual hull of the object; the maximum volume consistent with the silhouette appearance from a series of angles. Backprojecting the silhouette from each plane occurs through intersecting solid visual cones. The comparison of points across two images is the basis of stereovision. These contour control vertices of similarity can be utilized to analyze concave regions captured in the projection contour. A major computational method for reconstructing "solid" objects is to use multiple images that consist of a series of 2D "slices" of the target. By adjusting the slice depth of field and the focus plane, volumetric rendering can be partaken to construct a 3D model from the edges of the 2D slices. In contrast, electron tomography reconstructs 3D structures from projections taken at regular tilt intervals. The central section of the Fourier Transform (FT) is a thin slice of the 2D FT of the orthonormal projection image. If the FT is extended toward infinity in length, the central slice in real space collapses to a single line. These slices (the line) can be used to determine the alignment of the images by determining the angle of the slices that match. These slices, with appropriate filtering to limit the oversampled low frequency contribution, can also be used to reconstruct the base image by creating a new FT from the central slices based on alignment angles. By having many thousands of such projections, a suitable reconstruction can be created.

However, the problem of reconstruction is different when one has only a single orthographic projection image from which to derive a 3D model, as opposed to having many projections at different angles. The Fourier slice approach cannot be applied when there is only a single image for reference. In this research, we explore another option for deriving a 3D reconstruction of a particle from a single projection image, one that uses the hidden Markov model (HMM) to find the most likely 3D structure that could have produced the observed image.

The hidden Markov model is a tool for modeling a discrete finite automaton where the next state of the system is based only on the current state, and for each state the system emits an observable output. The states themselves that cause the output cannot be observed. Given a matrix representing the probabilities that the system transitions from one given state to another, a matrix representing the probabilities that a state emits an output, and a vector representing the probabilities of the initial state of the system, we are able to answer questions about the system such as

1. What is the most likely sequence of states that generated the output?
2. What is the most likely state at a given position in the sequence?

Both problems can be solved efficiently using the dynamic programming algorithms known as the Viterbi algorithm, which solves problem 1, and the forward-backward algorithm, which solves problem 2. In this research, we consider the state of a 3D volume, representing an LDL particle in this case, as being in one of finitely many states.

A state transition is the transition between one 3D voxel to another in a given direction. An observation is the density value of the orthographic projection image garnered via Cryo-TEM imaging. We calculate state transition and state to observation matrices from a 3D model representing the particle, train the matrices with projection images generated from the model, and use that statistical data to compute the most likely series of voxels in the genuine 3D representation of the particle, using the methods described above.

In addition to solving these problems, you can also train the original model by adjusting the state transition matrix and state to observation matrix based on new data in a process called "training," which can be performed efficiently using the Baum-Welch algorithm

The method of HMM 3D reconstruction used in this project is not only applicable to the lipoprotein, but to any particle or 3D volume. Applying the learning property of the HMM training procedure facilitates a more accurate 3D structure reconstruction from a single image than what was possible through conventional means.

3. Definitions, acronyms, and abbreviations

- Low-density lipoprotein (LDL): LDL cholesterol is often referred to as "bad cholesterol" because too much is unhealthy and easy to form plaque and hard deposit leading to arteries clog and make the artery less flexible. Approximately 50 percent of the weight of an LDL particle is cholesterol and only 25 percent is protein. Heart attack or stroke will happen if the hard deposit blocks a narrowed artery.

- High-density lipoprotein (HDL): HDL is often referred to as "good cholesterol". It contains 20 percent cholesterol by weight and 50 percent protein. It acts as a scavenger to help to remove LDL from the artery back to the liver.

- Cryogenic transmission electron microscopy (Cryo-TEM): a mainstream technology for studying the architecture, size, shape and internal structure of cells, viruses and protein assemblies at molecular resolution.

- Nanoparticles are particles with at least one dimension, between 10 and 100 nanometers in size. Nanoparticles are effectively a bridge between atomic or molecular structures and bulk materials. They are widely used on drug delivery and drug modification.

- Hidden Markov Model (HMM) is a tool for modeling stochastic Bayesian processes with hidden states, where the probability distribution of transitioning to the next state is known and is dependent only on the current state.

- Voxel: A 3D unit of an image. It is the 3D version of a pixel.

4. References

1. Good vs Bad Cholesterol. American Heart Association. http://www.heart.org/HEARTORG/
Conditions/Cholesterol/AboutCholesterol/GoodvsBadCholesterol_UCM_305561_Article.jsp
- .VuhTsmQrIy4.
2. Otvos, J., Marabini, R., Masegosa, I., Martın, M. S., Marco, S., Fernandez, . J., de la Fraga,
L. G., Vaquerizo, C., and Carazo, J. Radio signals give new spectrum for cholesterol lipoprotein
readings. In Arteriosclerosis, Thrombosis and Vascular Biology: Journal of the American Heart
Association (July 1998), pp. 2223–2231.
3. Jeyarajah, E. J., Cromwell, W. C., and Otvos, J. D. Lipoprotein particle analysis by nuclear
magnetic resonance spectroscopy. In Clinics in Laboratory Medicine (Dec. 2006), pp. 847–870.
4. Ewen Callaway. "The revolution will not be crystallized: a new method sweeps through
structural biology". Nature 525: 172–174. doi:10.1038/525172a. (9 September 2015)

2. Problem definition

Non-uniform transparent objects, such as LDL can be imaged by cryogenic electron microscopy.
During this processing, the target particles are flash-frozen from their natural state so that structure
could be preserved. During the image process, a single particle could be imaged at relatively large
or small rotation angles. The target particle was identified and imaged at multiple rotation angles
to acquire set of images, but during the imaging process, the sample may need to be relocated in
order to preserve the focus and keep the target(s) within the focus region of the cryo-TEM.

If the rotation is made in a small angle, the cryo-TEM sample holder is rotated through a limited
small range of angles each time. Therefore, refocusing may not be needed and this will reduce the
damage to the biological sample by the electron beam during the focus process. However, if the
geometric parameters of the particles are needed, the small rotation angles will limits the amount
of the information that may be extracted from the orthographic projection. This conflicts leads to a
natural tradeoff between sample integrity and processing requirements.

Images can be acquired with the small-angle rotations and software program can be applied in
order to obtain an improved mapping between the rotated 2-D projection images and geometric
parameters of a 3-D model. This mapping can be finalized for the extraction of pertinent geometric
information that can lead to an understanding of the internal structure of the interested particle. An
obvious desire and most important idea results from this process is that a single image of the
particle is used for reconstruction, then the refocusing is eliminated and varying beam damage
between images is eliminated. Adding data analysis and add the data into a image database, the
machine learn process can be built into the program, and the ultimate aim is that the 3D
reconstruction of the particle can be achieved by simple, single, individual image.

2.1 List of accomplishments

We have a renderer that will render images based on the density of a projection.
We have a means of rendering images in bulk.

We have derived a general formula for an A matrix at arbitrary rotation based on orthogonal A matrices in higher dimensional space. In 2D space, an A matrix at an arbitrary rotation is given by

$$A^\theta = (A^X)^{|\cos\theta|}(A^Y)^{|\sin\theta|}$$

When $\cos(\theta)$ is negative, take the stochastic transpose of A^x, and when $\sin(\theta)$ is negative, take the stochastic transpose of A^y. That is, take the transpose of the matrix, and multiply the entry of each matrix by the column sum of the cumulative transition matrix corresponding to that entry, divided by the row sum of the cumulative transition matrix for that entry.

This formula can be extended to 3D A matrices, and is given by:

$$A^{\theta,\phi} = (A^X)^{|\cos\phi\cos\theta|}(A^Y)^{|\cos\phi\sin\theta|}(A^Z)^{|\sin\phi|}$$

following the same process when any of the exponents within the absolute value bars are negative, taking the stochastic transpose of the matrix. The angles φ and θ correspond to the angles used in the spherical coordinate system, where φ is the angle between the z projection and the xy plane, and θ is the angle between the x and y projections. The A matrix we calculate is used when calculating the probability distribution in the 3D variant of the Viterbi algorithm we have derived:

$$V_{r,s,t,k} = P(y_{r,s,t}|k) \cdot$$
$$\max_{x,y,z\in S}\left\{a^X_{x,k}\cdot V_{r-1,s,t,x}\cdot a^Y_{y,k}\cdot V_{r,s-1,t,y}\cdot a^Z_{z,k}\cdot V_{r,s,t-1,z}\right\}$$

Maximizing this algorithm in one pass requires some extra memory usage to cache the result of the function for each input, but the result is a far lower time complexity, on the polynomial order. The P term can be calculated as it is a function of the sum of the optimal states in the row and column being processed, and it is our theory that this could supersede the need for a B matrix for this problem, which via Lewis's method was calculated anyway, via artificial 2D projections of the initial guess model. Instead, the A matrix is trained using new data and the B matrix is a function of the A matrix rotated at the observed angle, the previous states calculated in the row/column, and the amount of states remaining to calculate. Since these states are cached under this implementation, it makes sense to use all the available information to construct a better model instead of using a strictly observation-independent model. We expect that the result will be a very accurate reconstruction.

The software for reconstruction which we have produced thus far is available at:

It is capable of rendering a PDB file as a 3D volume as a stack of 2D slices in PNG format, and calculating the three orthogonal A matrices for each axis. It is also capable of performing the 3D Viterbi algorithm on a set of data and providing a best fit reconstruction.

2.2 Problems encountered

A decent amount of our time involved trying to find the correct program in order to produce the desired filetype (.dcm) from the filetype we intended to create (.obj). By the time we found a way to produce files that could be loaded into ImageJ, we abandoned the idea of using .dcm for .png files.

Rendering PDB files as OBJ in VMD did not include the atomic indicative textures, therefore the variations in electron density intramolecularly and intermolecularly were not able to be put into grayscale form, therefore the final Maya rendering was obsolete for rendering of the model.

We encountered the same problems as Lewis in his 3D reconstruction method, which included a lack in the ability of the algorithm to solve a multidimensional system. We encountered issues creating an efficient implementation of the higher dimensional Viterbi algorithm, but these issues were solved using an efficient caching strategy.

2.3 Team member contribution

Raymond Pulver: Developed the theory of higher dimensional hidden Markov models, developed the 3D Viterbi algorithm, developed an algorithm for the rotation of a set of linearly independent A matrices.

Neal Buxton: Edited the ImageJ Volume Viewer to render images based on intensity. Created a Macro in ImageJ that allows batch rendering of the images, adding noise and salt and pepper.

Xiaodong Wang: Searched biological literature that related to this project. Organized and generated the draft of proposal and final report for other team member to add on the content. Provided biological suggestion to the computer team part.

John Lucci: I compiled an LDL particle with a more diverse and accurate lipid profile than has been previously analyzed in 3D form. This required 3D Surf rendering of two lipids not previously characterized as 3D structures searchable in the RCSB database, KOIDA-PC and POV-PC. Most strenuously, I made a de novo 3D structural model of ApoB, only 1000 out of 5500 residues of which were able to be characterized by a homology model. The structure of the other 4500 residues were formed through ab initio synthesis.

2.4 Implementation status

The routines that perform the Viterbi algorithm are complete with the exception of a probability distribution that describes the likelihood of a given observation, based on what is currently known about the model at that stage of reconstruction. Once this function is implemented the reconstruction will be able to be performed with exceptional speed compared to the previous methods used, with far greater accuracy.

In short, our programs are almost complete. The ImageJ Volume Viewer seems to now be working as intended as it can be used for batch processing as well as for displaying volume data. The macro is functional, but the graphical interface could still use some work.

2.5 Words of wisdom to anyone using this work

A higher dimensional model is necessary for a reasonable reconstruction. Future researchers exploring this problem should learn from Lewis Collier's results that the one dimensional model is not suitable for this problem. If the multidimensional algorithm developed by our team is employed, it is necessary to employ a caching strategy to avoid recalculating any points in the path while traversing the volume.

On another note, images take up a lot of memory and should not be held in memory any longer than absolutely necessary. With that in mind, it is best to take the time to render them separately when dealing with large numbers of images. Arrays/stacks will cause memory overflow once a sizeable number of images are processed.

3. Results

In this project report, we introduce a new approach to reconstruct of the particle object. The basis imaging modality is, but not limited to the Cryo-TEM. Huge amount of images are required for traditional existing methods of reconstructing biological particles. Also, significant computational process need to average and reduced the noise to produce a reconstruction that represents the ensemble average particle. The approach presented in this report requires only one single individual image and a reasonable computational engine to provide a true individual particle reconstruction. This method uses an a priori and more efficient model from biological and/or physical constraints to deduce the voxel to voxel statistics, and thus, to determine the most probable voxel states. Comparing to the traditional techniques, our approach provides reconstruction of multiple shapes from a single individual image. A further promising aspect of this technique is that the detailed internal structure of the interested particle can also be gleaned that can lead to a better understanding of the model. Our reported approach offers a much more enhanced but simpler approach to solving the transparent object reconstruction problem than required by existing traditional methods.

4. Future work and conclusions

Even though some encouraging results shown in this research so far, there are additional areas in which this model and program can be studied for further advancements and improvement. Spherical objects are similar models. Any other new shapes can be processed and added into the database for machine-learn before a new image is studied. The resolution differences between the actual and modeled shape still need to be improved in the future.

The program presented in this report provides its effectiveness due to its innovative construction. This approach offers a very simple reconstruction method that requires significantly less computational processing compared to standard methods used nowadays. The construction allows for massive data analysis at the same time where each chain, or small groups of chains, can be computed concurrently. Therefore, our program only requires minutes of processing on a modest laptop computer versus many hours of computations by using heavy-duty workstations for the existing models. Only one individual image is used during the image analysis process, and the noise statistics can be saved, thus it has the ability to work regardless of the noise found in non-averaged images. Also, eliminating errors arising from the multiple image analysis process allows the inner structure of the interested particle to be reconstructed more accurately.

5. Appendix (attach the program and all table flows)

ImageJ Macro:

```
setBatchMode(true);
run("Image Sequence...", "convert_to_rgb sort use");
end_dir=getDirectory("Choose a Directory");
stack1 = getImageID;
stack2 = 0;
angles=newArray(0,90,180,270);
devs=newArray(-5,0,5);
g_n=newArray(0,25,50);
for (fn=0;fn<g_n.length;fn++){
        dname=end_dir+"gsn_"+g_n[fn];
        if (File.isDirectory(dname)==0) {
        File.makeDirectory(dname)
        }
        dname_plain=dname+"\\plain";
        if (File.isDirectory(dname_plain)==0) {
        File.makeDirectory(dname_plain)
        }
        dname_sandp=dname+"\\sandp";
        if (File.isDirectory(dname_sandp)==0) {
        File.makeDirectory(dname_sandp)
        }
    }
```

```
selectImage(stack1);
n = angles.length;
ay=0;
dy=0;
for (i=0; i<n; i++) {
        showProgress(i, n);
        ax=angles[i];
        for (k=0;k<n;k++){
        az=angles[k];
        for (x=0;x<devs.length;x++){
        dx=devs[x];
        for (z=0;z<devs.length;z++){
        dz=devs[z];
        selectImage(stack1);
        run("Volume Viewer", "display_mode=5 scale=2.5 axes=0 interpolation=4
angle_x="+(360+ax+dx)%360 +" angle_y="+((360+ay+dy)%360)+" angle_z="+(360+(az+dz)
%360)+" bg_r=255 bg_g=255 bg_b=255" );
        run("Copy");
        w=getWidth; h=getHeight;
        close();
        thisname="Plots_"+((360+ax+dx)%360)+"_"+((360+az+dz)%360)+".png";
        newImage(thisname, "RGB", w, h, 1);
        run("Paste");
        run("8-bit");
        fnameplain=end_dir+"gsn_0\\plain\\"+thisname;
        for (fn=0;fn<g_n.length;fn++){
        run("8-bit");
        dname_plain=end_dir+"gsn_"+g_n[fn]+"\\plain";
        if (g_n[fn]>0){
        run("Add Specified Noise...", "standard="+g_n[fn]);
        }
        saveAs("PNG", ""+dname_plain+"\\"+thisname);
        dname_sandp=end_dir+"gsn_"+g_n[fn]+"\\sandp";
        run("Salt and Pepper");
        saveAs("PNG", ""+dname_sandp+"\\"+thisname);
        close();
        open(fnameplain);
        }
        close();
        }
        }
        }
}
run("Select None");
```
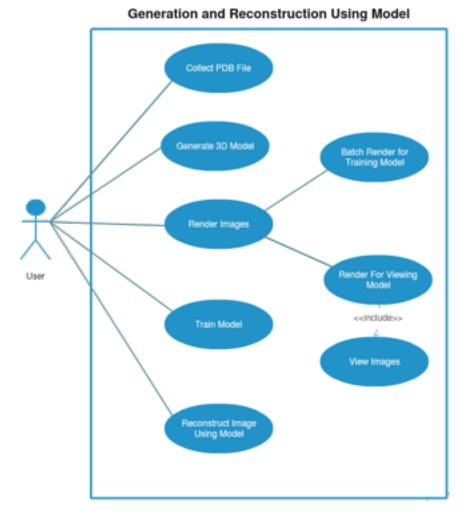
setBatchMode(false);

ImageJ Volume Viewer:

There are too many modules to put all of the code here.
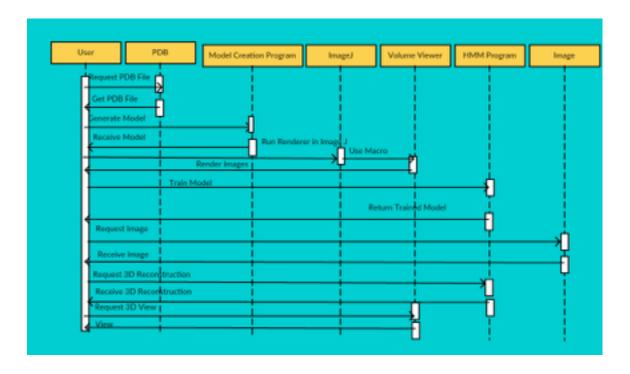I will list the edited modules and the changes made:

Control.java--edited to include Projection Sum and Current Pixel
Gui.java -- edited to include Projection Sum and Current Pixel
Interpolation.java -- edited to return whichever pixel the requested point is in
Pic.java -- Edited to return a weighted sum of pixel values:
1. Collect the longest possible projection within the image.
2. Collect the maximum value pixel.
3. Multiply the longest projection by the maximum value pixel.
4. For each projection sum all values and divide by the value calculated in step 3.

Volume_Viewer.java -- Edited for batch processing and to get limits and max value of Image

Generation and Reconstruction Using Model

6. The code and all written docs uploads to Sakai
ImageJ VolumeViewer and macro: https://github.com/NealBuxton/A3D-SRINP
Voxelizer and 3D HMM solver: https://github.com/raypulver/viterbi