Study of amyloid aggregation in Alzheimer's Disease using clustering analysis of scanning x-ray microdiffraction patterns

Yan Zhang[1] and Lee Makowski[2,3]
11/29/2016

Abstract— **Scanning X-ray microdiffraction (SXMD) is used to study the amyloid plaque associated with Alzheimer's Disease (AD) in molecular length-scale. It has been shown that the dissemination of amyloid undergoes the progression of AD. SXMD using micro size beam at Advance Photon Source, Argonne National Laboratory is capable to collect 2500 diffraction patterns in a 250x250 micro field of view of AD sample provided by clinical researchers from the Massachusetts Alzheimer's Disease Research Center. Here, in order to analyze these large amount of diffraction data, we utilized and compared different clustering methods and showed promising results. The clustering, correlation and regression results clearly showed the distribution of molecular constituents and structural heterogenity of amyloid fibrils and plaques as well as other tissues. This study opens a new perspective of AD research connecting clinical experiments, imaging informatics and data analysis.**

I. INTRODUCTION

Alzheimer's Disease (AD) is a fatal, neurodegenerative disease common in elderly populations. Many senior people have suffered AD worldwidely, which has huge impact on both healthcare system and medical economy. Many clicical and pharmaceutical research has been performed to understand the cause of this disease but no prevention or treatment strategies been made successful: The molecular mechanism underlying AD has not been understood. Accumulaiton of amyloid has been shown to result in formation of plaques in brain tissue [1]. X-ray scattering as well as other imaging techniques have been used to study amyloid in AD for decades [2][3][4]. Traditional X-ray scattering techniques used beam size of hundreds of micrometers. Modern X-ray technology at synchrotron facilities may utilize micrometer size beam - a scanning X-ray microdiffraction - possible for researchers to examine the distribution of molecular structures in their tissue sample, which is an ideal tool to study amyloid aggregation in AD [5]. SXMD can collect hundreds to thousands of diffraction patterns in thirty minutes by x-ray scanning the different positions of the tissue in a certain area [6]. The amyloid aggregation may only be captured in a limited number of diffraction patterns over thousands of images recorded. In this section, we used machine learning methods to analyze thousands of SXMD data of brain tissue from AD patients. The analysis provided very useful information for biomedical or biological researchers and proved to be efficient and accurate.

II. SAMPLE PREPARISION AND DATA COLLECTION

A. Sample preparation from Alzheimer's patients

Brain tissue samples are provided by Massachusetts Alzheimer's Disease Research Center (MADRC) at Massachusetts General Hospital (MGH). Tissue is collected and fixed in 10% formalin. Representative blocks are reomoved; treated with 99% formic acid for an hour; put through isoproply alcohol, then xylene and finally paraffin wax. The information about brain sample donor is limited and anonymized.

B. SXMD data collection

The SXMD data collection experiment was carried out by collaborators at Advanced Photon Source 23IDB at Argonne National Laboratory [7]. Scattering data were collected over a range from 0.008 $\mathring{A}^{-1}$ to 0.35 $\mathring{A}^{-1}$ in reciprocol space. Brain tissue are exposed to a 1 micrometer size X-ray beam which shots and scans the sample in a 50x50 grid, see figure 1. It took a second to capture each diffraction pattern and roughly 42 minutes were used to collecte these total 2500 images. The intensities of 2-D diffraction patterns are circular averaged and normalized in order to get 1-D intensity data to reduce the computational complexity. The optical micrograph and 50x50 grid of microdiffraction are shown in figure 1.

III. INTENSITY DATA ANALYSIS

A. Weighted k-means clustering

The AD brain tissue scattering is more likely a 'protein in solution' scenario and intensities were smooth exhibiting only subtle differences among 2500 patterns. We use weighted k-means clustering which the distance metric is weighted by a Gaussian fuction. The Gaussian weighting funtion is:

$$p(q|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(q-\mu)^2}{2\sigma^2}} \tag{1}$$

where $\mu = 0(1/\AA)$ for SAXS weighting and $\mu = 0.23(1/\AA)$ for WAXS weighting. $\sigma = 0.15(1/\AA)$ is a fixed value. The k intensity centroids are $\bar{I}_1(q), ...\bar{I}_k(q)$. The two-steps are:

*This work was supported as part of the Center for Direct Catalytic Conversion of Biomass to Biofuels (C3Bio), an Energy Frontier Research Center funded by the U.S. Department of Energy, Office of Science, Basic Energy Science under award #DE-SC0000997.

[1]Department of Electrical and Computer Engineering, Northeastern University, Boston MA yzhang@ece.neu.edu

[2]Departments of Bioengineering; [3]Department of Chemistry and Chemical Biology, Northeastern University, Boston MA l.makowski@neu.edu

Step 1: assign intensity data to clusters by comparing the distance to centroids:

$$C_i^{(t)} = \{I_j(q) : argmin\{\sum_{q=0}^{n-1} p(q|\mu,\sigma)|I_j(q) - \bar{I}_i(q)^{(t)}|^2\}\} \quad (2)$$

and $i \in \{1, ..., k\}$.

Step 2: update new centroids for t+1 step:
the number of k can be chosen by the knee point of Singular value decomposition (SVD) of intensity data. The circularly averaged intensities of each diffraction pattern are stored in a matrix **I**. SVD algorithm decomposes the matrix I into three matrices **U**, **S** and **V**:

$$\mathbf{I} = \mathbf{USV}^T \quad (3)$$

where I is a M x N matrix, M represents the length of scattering angle and N is the number of circularly averaged intensities (2500 diffraction patterns). U is a M x M matrix and V is a N x N matrix whose columns are orthonormal basis vectors. S is a M x N diagonal matrix and the descending entries are the singular values of matrix **I**. The first k singular values before 'elbow' point in the the singular values plot means the number of basis vector needed to reconstruct matrix **I**, which potentially indicates the number of clusters we need to partition the 2500 intensities.

B. Spectral clustering method

In order to find a reasonable number of clusters needed to reduce the dimensionality while maintaining the features of the intensity dataset at same time, we adapted spectral clustering method which has been shown to work well on both spherical and elliptical data [8]. The 50x50 intensities are formed in serial in an array $\mathbf{I} = \{I_1(q), I_2(q)...I_n(q)\}$, where $n = 2500$. Each intensity $I_i(q)$ is a m dimensional vector, aka. $\mathbf{I} \in \mathbb{R}^m$.

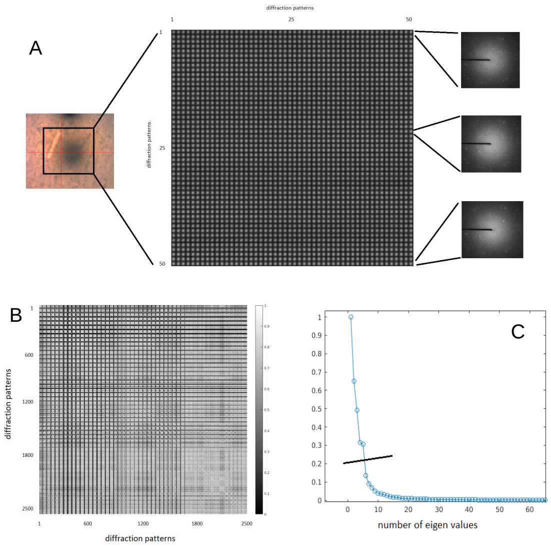

Fig. 1: (A). Optical microscope image of AD sample, the 50x50 SXMD patterns and three individual diffraction patterns. (B). Affinity matrix using all circular averaged intensities of AD. (C). Eigen values of matrix L to find the k-th largest eigen vectors for dimension reduction.

An affinity matrix A is formed:

$$A_{ij} = e^{-||I_i - I_j||^2}, \ if \ i \neq j \quad (4)$$

and $A_{ij} = 0$ for $i = j$.

Define D as a diagonal matrix whose diagonal values are:

$$D_{ii} = \sum_{j=0}^{n-1} A_{ij} \quad (5)$$

Construct matrix L that:

$$L = D^{\frac{1}{2}} A D^{\frac{1}{2}} \quad (6)$$

The eigenvalues and eigenvectors of L are computed and matrix X is formed by selecting the first k-th largest eigenvectors, in another word, X is a nxk dimension reduced matrix. Here, k is chosen by identification of the knee point of eigenvalues of L.

Normalized each row of matrix X to have a matrix Y whose row is in unit length:

$$Y_{ij} = \frac{X_{ij}}{(\sum_{j=0}^{k-1} X_{ij}^2)^{1/2}} \quad (7)$$

Each row in matrix Y is treated as a point in k dimensional space, $Y \in \mathbb{R}^k$ and k-means clustering method is used to partition the n points into k groups.

Here the k centroids are $\bar{Y}_1, ... \bar{Y}_k$. The two-steps are following:

Step 1: assign k dimensional points into clusters by comparing the distance to centroids:

$$C_i^{(t)} = \{Y_p : ||Y_p - \bar{Y}_i^{(t)}|| \leq ||Y_p - \bar{Y}_m^{(t)}||\} \quad (8)$$

and $m \in \{1, ..., k\}$.

Step 2: update new centroids for t+1 step:

$$\bar{Y}_i^{(t+1)} = \frac{\sum_{Y_p \in C_i^{(t)}} Y_p}{\sum_{Y_j \in C_i^{(t)}} 1} \quad (9)$$

In the last, the original intensity data $\mathbf{I} \in \mathbb{R}^m$ are assigned into the k clusters using the same clustering labels of $Y \in \mathbb{R}^k$.

C. Clustering using simulated amyloid molecular structure

Amyloid molecular structure (PDB:2lmp) is shown in figure . The simulated scattering intensity using Debye formula in cylindrical coordinates is given by [9]:

$$I_{sim}(R,Z) = \sum_j \sum_k f_j(R,Z) f_k(R,z) J_0(2\pi r_{jk} R) cos(2\pi z_{jk} Z) \quad (10)$$

with scattering vector R defined as:

$$R = 2sin(\theta)/\lambda \quad (11)$$

where $I_{sim}(R,Z)$ is the diffraction pattern of amyloid. R and Z are equatorial and meridional scattering vectors in reciprocal

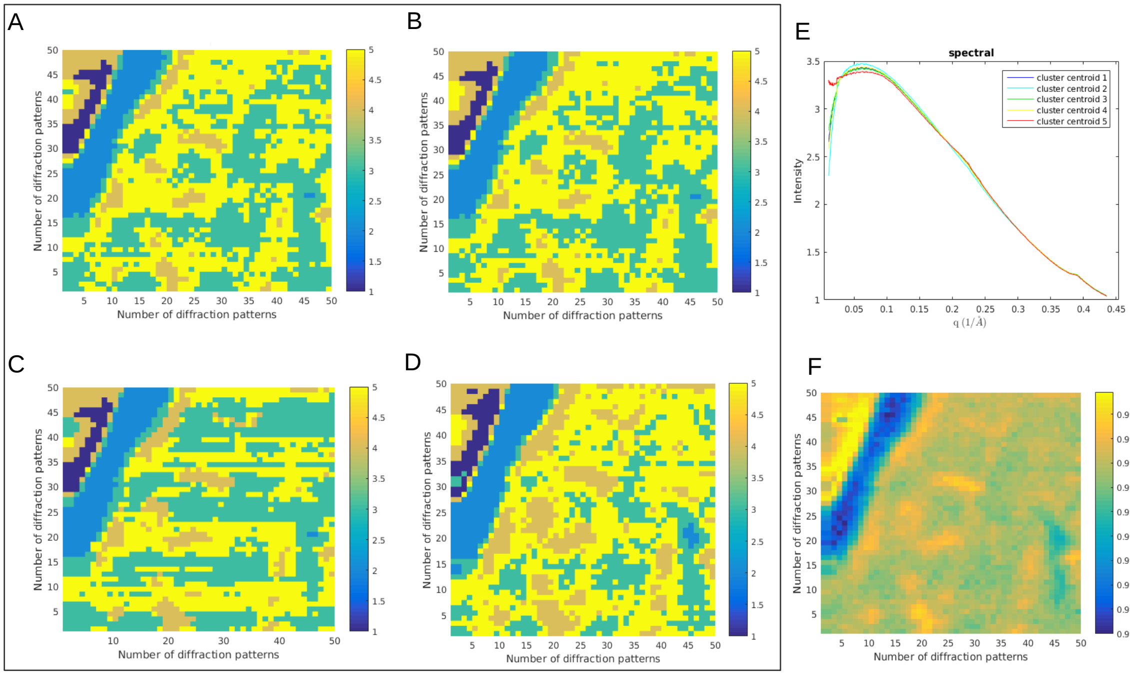

Fig. 2: (A). K-means clustering result. (B-C). Weighted k-means clusteing result on low-angle and high-angle intensity. (D). Spectral clustering result. (E). Cluster centroids from spectral clustering. (F). Correlation coefficient between centroid intensity of amyloid and all 2500 intensities.

space, λ is x-ray wavelength. The f_j and f_k are atomic form factors, r_{jk} and z_{jk} are the radial and axial components of the inter-atomic distance between two atoms in cylindrical coordinates and J_0 is the zero-order Bessel function of the first kind. A fast Debye formula was derived and used to simulate two-dimensional diffraction patterns [10].

We proposed that the amyloid plaue is an ensemble of many fibrous molecules in different orientation, thus we used a linear system to represent this multi-orientation model of amyloid aggregation in AD

$$I_{exp} = \sum_{i=0}^{N} \alpha_i I_{sim,i} + bgr + \varepsilon \tag{12}$$

and to find the minimization of

$$argmin_{\alpha} \frac{1}{2} || I_{exp} - \sum_{i=0}^{N} \alpha_i I_{sim,i} ||_2^2 \tag{13}$$

which is subjected to

$$\sum \alpha_i = 1, \quad 0 \leq \alpha_i \leq 1 \tag{14}$$

where I_{exp} is the circularly averaged experimental intensity of SXMD pattern, α_i and $I_{sim,i}$ are weighting coefficient and scattering intensity of θ_i degree from equatorial in the simulated diffraction pattern (θ_0 is equatorial intensity and θ_{90} is meridional intensity). ε and bgr is considered as the Gaussian noise and scattering background.

Last, a sigmoid function is used returning value S to partition SXMD patterns:

$$S = \frac{1}{1 + e^{(-\beta^{\mathbf{T}} \omega)}} \tag{15}$$

where $\beta = [\alpha_0, \alpha_1, ..., \alpha_{90}]^T$ and $\omega = [\frac{1}{\theta_0+\varepsilon}, \frac{1}{\theta_1}, ..., \frac{1}{\theta_{90}}]^T$.

IV. Result and Discussion

The affinity matrix and eigen values of these 2500 diffraction patterns were shown in figure 1 B and C. The first five singular values were chosen to reconstruct the dataset in lower dimension of intensity data. The 2500 diffraction patterns were analyzed using regular kmeans, weighted kmeans and spectral clustering methods, shown in figure 2 A-D. The centroids of each clusters were plotted in figure 2 E. The most significant cluster centroid (red) can be easily picked up and the correlation coefficient between itself and intensities of all difffraction patterns are shown in figure 2 F. Comparing the clustering results with optical micrograph, the diffraction patterns from small blood vessel on the top left could be easily identified (light blue). A dense region of amyloid is located on the upper left of the blood vessel (dark blue) in figure 2 D. Correlation coefficient map between centroid of amyloid intensity and all 2500 intensities showed several other regions of amyloid in figure 2 F, (yellow). Through the clustering analysis, we can easily classify the diffraction patterns of dense amyloid, weak amyloid, empty blood vessel and other scattering contents.

Clustering analysis helps to partition these SXMD data without having information about amyloid structure. Simulated diffraction pattern of amyloid fibrous molecule can further assists us on structure characterization by comparing intensities in reciprocal space. Since the SXMD images were circularly average, the orientation of amyloid molecules played an important role contributing different scattering intensities. The multi-orientation model fitting to the SXMD intensities can estimate the weighting factors of each intensity trace of angular shift from equatorial plane. The diffraction pattern of much aggregated amyloid will apear to have higher weighting factors on equatorial intensity than others.

Processing and analysis of large image data becomes an emerging challege in biomedical research. In the Alzheimer's Disease study in molecular level, scanning x-ray microdiffraction techniques were adapted using the world-leading synchrotron source to generate huge amount of data per sample. The screening and selection of amyloid associated intensity data from these 2500 diffraction patterns were done by using clustering methods. The clustering results provide preliminary information about these data for biomedical and biological research.

References

[1] Thal, Dietmar R., et al. "Phases of A?-deposition in the human brain and its relevance for the development of AD." Neurology 58.12 (2002): 1791-1800.

[2] Eanes, E. D., and G. G. Glenner. "X-ray diffraction studies on amyloid filaments." Journal of Histochemistry and Cytochemistry 16.11 (1968): 673-677.

[3] Kirschner, Daniel A., Carmela Abraham, and Dennis J. Selkoe. "X-ray diffraction from intraneuronal paired helical filaments and extraneuronal amyloid fibers in Alzheimer disease indicates cross-beta conformation." Proceedings of the National Academy of Sciences 83.2 (1986): 503-507.

[4] Petkova, Aneta T., et al. "A structural model for Alzheimer's beta-amyloid fibrils based on experimental constraints from solid state NMR." Proceedings of the National Academy of Sciences 99.26 (2002): 16742-16747.

[5] Liu, Jiliang, et al. "Amyloid structure exhibits polymorphism on multiple length scales in human brain tissue." Scientific Reports 6 (2016).

[6] Zhang, Yan, Jiliang Liu, and Lee Makowski. "A new pre-processing method for scanning X-ray microdiffraction patterns." Biomedical Circuits and Systems Conference (BioCAS), 2015 IEEE. IEEE, 2015.

[7] Fischetti, Robert F., et al. "Mini-beam collimator enables microcrystallography experiments on standard beamlines." Journal of synchrotron radiation 16.2 (2009): 217-225.

[8] Ng, Andrew Y., Michael I. Jordan, and Yair Weiss. "On spectral clustering: Analysis and an algorithm." Advances in neural information processing systems 2 (2002): 849-856.

[9] Inouye, Hideyo, Paul E. Fraser, and Daniel A. Kirschner. "Structure of beta-crystallite assemblies formed by Alzheimer beta-amyloid protein analogues: analysis by x-ray diffraction." Biophysical journal 64.2 (1993): 502.

[10] Zhang, Yan, et al. "Diffraction pattern simulation of cellulose fibrils using distributed and quantized pair distances." Journal of Applied Crystallography 49.6 (2016).

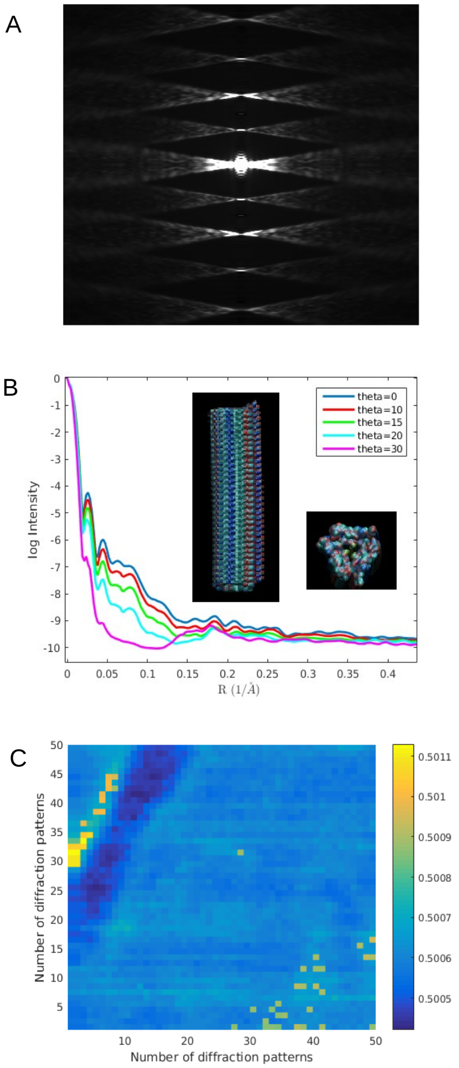

Fig. 3: (A). Simulated diffraction pattern of amyloid molecule. (B). Intensities in various orientation, cross-section and longitudinal view of amyloid molecule. (C). Clustering result using multi-orientation model.