Calculation of pair-distance distribution function using a pre-trained codebook

Yan Zhang,[a] Michael Crowley,[b] Jacob Hinkle[c] and Lee Makowski [d,e]*

[a]*Material Science Division, Argonne National Laboratory, Lemont, IL 60439 USA,* [b]*Biosciences Center, National Renewable Energy Laboratory, Golden, CO 80401, USA,* [c]*Computational Science Center, National Renewable Energy Laboratory, Golden, CO 80401, USA,* [d]*Department of Bioengineering, Northeastern University, Boston, MA 02148, USA, and* [e]*Department of Chemistry and Chemical Biology, Northeastern University, Boston, MA 02148, USA. E-mail: yz@anl.gov*

Abstract

X-ray scattering is an important tools to understand the deconstruction of cellulose microfibrils due to chemical pretreatment. Application of this tool to biomass undergoing degradation is challenging because of the heterogeneous nature of the fibrils which are being damaged after chemical deconstruction. In this paper, we describe a fast and accurate method for utilizing scattering intensities to determine the pair-distance distribution of different fibrillar structures. A set of cellulose microfibrils of different shape, size and twist are generated, and are used to calculate a library of pair-distance distributions and scattering intensities using Debye formula. A codebook is trained and generated as the inverse Fourier Transform operator for scattering intensities. A relative proportion of pair-distance distribution is the operation of the codebook on

the observed intensity. This method is shown fast and accurate with a small NRMSE between calculated and ground truth pair-distance distribution.

1. Introduction

Lignocellulosic biomass is an ideal candidate to be a sustainable and renewable bioenergy source producing ethanol and/or other fuels and chemicals (Lynd et al., 1991). The breakdown of biomass into cellulose fibrils, individual chains (Himmel et al., 2007; Himmel, 2009) and molecular fragments (Mok & Antal, 1992; Inouye et al., 2014) has been studied and X-ray diffraction is a key technique for monitoring the breakdown process and assessing the efficiency of various deconstruction techniques.

X-ray scattering collected at synchrotron sources have been used to estimate average cross-section size of cellulose fibrils (Fernandes et al., 2011; Thomas et al., 2013; Inouye et al., 2014; Zhang et al., 2015). It will greatly help researcher to understand the cellulose structural deconstruction if one can easily calculate pair-distance distribution function from synchrotron collected intensity data, so as to understand the relative abundance of different structures, which is naturally occuring from deconstruction process.

Here, we describe a process that will result in the calculation of pair-distance distribution of different fibril structures from x-ray scattering intensity. In order to do this, a large dataset of cellulose microfibrils including structures as small as 4-chain to as big as 81-chain are constructed with different cross-sectional shapes. Twisted version of these structures are also generated using molecular dynamics. From these fibril structures, a library of pair-distance distributions and corresponding intensities are calculated. A codebook mapping the intensities to the corresponding pair-distance distributions is then generated.

We demonstrate that we can use codebook as an operator on the observed intensities

to calculate relative distributions of different fibrillar structures and how pair-distance distributions are essential in this calculation.

2. Generation of cellulose microfibrils of different shapes

Charmm (Chemistry at HARvard Molecular Mechanics) program (Brooks et al., 2009) was used to simulate many hypothetical cellulose fibrous structures. The cellulose structure simulation was performed at the National Renewable Energy Laboratory (NREL) using DOMDEC fast parallel method (Hynninen et al., 2014). Charmm was first used to generate an 81-chain (9x9) diamond shape cellulose fibril which is 200 $\mathring{A}$ long (20 cellobiose units) and in cellulose Iβ crystalline structure (Langan et al., 2005). The structure template was trimmed and reshaped into different cellulose fibril shapes and lengths. These structures were subject to a process of structural dynamics in water buffer until fibrils equilibrated (in $\sim$ 800 ns). The library of simulated cellulose fibril structure includes 2x2-, 3x3-, 4x4-, 5x5-, 6x6-, 7x7-, 8x8- and 9x9-chain diamond shape, and 36-chain hexagonal shape as well as a 24-chains brick shape. In this library, each shape contains 1 untwisted and 99 twisted structures, the cross-section view of cellulose microfibrils, pair-distance distribution and intensities of this library are shown in figure 1.

3. Simulated intensity of cellulose microfibrils

A fast Debye method was developed and used to accelerate the intensity simulation process (Zhang et al., 2016). This method first calculates all interatomic distances and labels them by atom type into six groups, aka. C-C, H-H, O-O, C-H, H-O, C-O. Then, it uses a histogram to compress the large number of interatomic distances into a small number of discrete distances. The resulting interatomic distance arrays of atom type a and b is $\hat{r_{ab}}$ associated with population-distributions w_{ab}. Since the pair-distances are

atom type indexed, the atomic scattering factors can be moved out of the summation in the Debye formula, which becomes:

$$I_{ab}(R) = f_a(R) f_b(R) \sum_i w_{ab,i} J_0(2\pi r_{ab,i}^{\hat{}} R) \tag{1}$$

The final scattering intensity is the sum of six sub-intensities simulated by each atom type indexed pair-distance arrays.

$$I(R) = \sum_{ab \in S} I_{ab}(R) \tag{2}$$

where S=C-C, H-H, O-O, C-H, H-O, C-O. f_a and f_b are atomic scattering factors for atom a and b, and J_0 is zero-order Bessel function of first kind. The pair-distance distribution function $P(r)$ is the summation of the six population-distributions w_{ab}:

$$P(r) = \sum_{ab \in S} w_{ab} \tag{3}$$

4. Codebook for <pair-distance : intensity> transformation

The pair-distance distribution is conceptually the inverse Fourier Transform of the scattering intensity of cellulose fibril:

$$P(r) = \mathcal{F}^{-1} \{I(R)\} \tag{4}$$

Instead of calculating $P(r)$ from cylindrical averaged scattering intensity, we propose a method using a linear operator to replace the inverse Fourier Transform operation based on the linearity of Fourier Transform (Bracewell et al., 1986), by solving the linear system equation:

$$\mathbf{P} = \mathbf{IT} + \epsilon \tag{5}$$

Where $\mathbf{P} \in \mathbb{R}^{\mathrm{MxN}}$, $\mathbf{I} \in \mathbb{R}^{\mathrm{MxK}}$ and $\mathbf{T} \in \mathbb{R}^{\mathrm{KxN}}$. Rows of $\mathbf{P}$ and $\mathbf{I}$ are pair-distance distribution and scattering intensity profiles respectively. N and K are corresponding to d-range and 1/d-range in space and reciprocal domain. The transformation matrix

$\mathbf{T}$ is the codebook and ϵ is additive Poisson noise. Given a set of simulated intensities and pair-distance distributions, we need to calculate the estimated $\hat{\mathbf{T}}$ by:

$$argmin_{\mathbf{T}} \left\{ \frac{1}{2} ||\mathbf{P} - \mathbf{IT}||^{\mathbf{2}} \right\} \tag{6}$$

or use the least-square method:

$$\hat{\mathbf{T}} = (\mathbf{I^T I})^{-\mathbf{1}} \mathbf{I^T P} \tag{7}$$

5. $P(r)$ calculation from input intensities

For any given scattering intensity I_s, the estimated pair-distance distribution $\hat{P}_s(r)$ can be reconstructed using the estimated codebook:

$$\hat{\mathbf{P}}_{\mathbf{s}} = \mathbf{I_s} \hat{\mathbf{T}} \tag{8}$$

Here, we used 4-, 9-, 16-, 36-, 49-, 64- and 81-chain diamond intensities and pair-distance distributions as training data and 24-chain brick, 25-chain diamond and 36-chain hexagonal as test data. In this setting, the training data includes only diamond shape but in many different size and the test dataset has three different types of shape and size, which is considered as unbiased. Each shape in the training dataset has 100 structures and in total 700 pair-distance distributions and scattering intensities for training. The ground-truth and estimated pair-distance distributions for test data are shown in figure 2. We can see that the estimated and ground truth pair-distance distributions are very similar. The NRMSE ($\frac{RMSE}{P_{s,max} - P_{s,min}}$) of the three estimation to their ground-truth are: 0.01, 0.05 and 0.02.

6. Discussion

In this paper, we demonstrated a pair-distance distribution estimation method using input intensities and pre-trained codebook, which is shown easy and accurate. We first

generated cellulose microfibrils in different shapes and twist using CHARMM program. A library of pair-distance distributions and scattering intensities of these structures were simulated and a codebook was calculated. The library used to generate the codebook consisted microfibrils in different size but same basic shape. Although the test data had microfibrils in different shape and size, the reconstruction of $P_s(r)$ still showed accuracy. This indicates that the training dataset has basis vectors that are complete for pair-distance distribution reconstruction of any shape/size larger than 4-chain and smaller than 81-chain (dimension of the codebook). This codebook is shown to be useful for Synchrotron data collection and analysis.

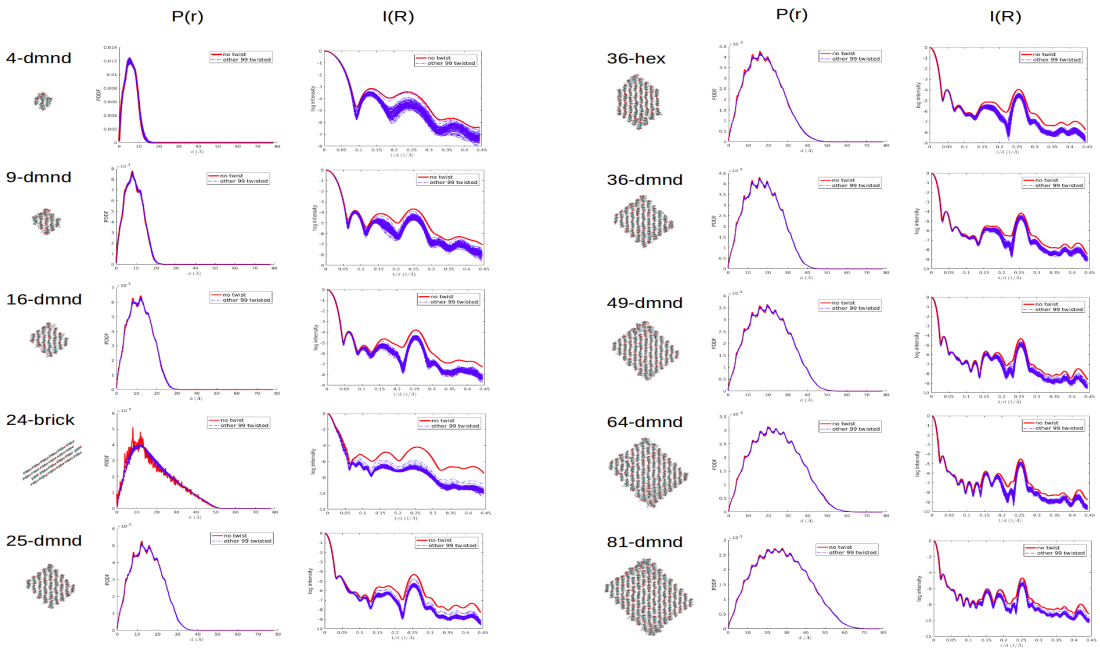

Fig. 1. A library of cellulose microfibrils (structures, pair-distance distributions and scattering intensities) from left to right. The library includes 10 shapes and 100 structures for each shape.

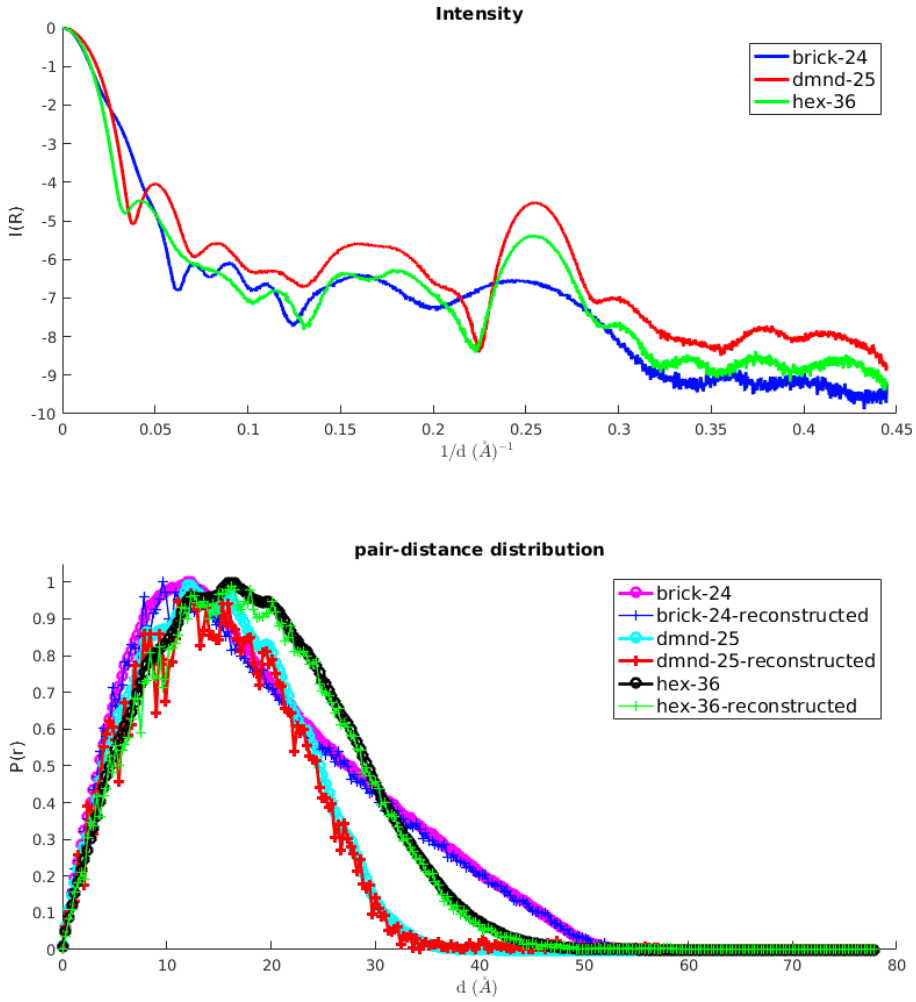

Fig. 2. Simulated Intensities (Upper) and Estimated/Ground-Truth (Lower) pair-distance distribution of 24-chain brick, 25-chain diamond and 36-chain hexagonal structures of cellulose microfibrils.

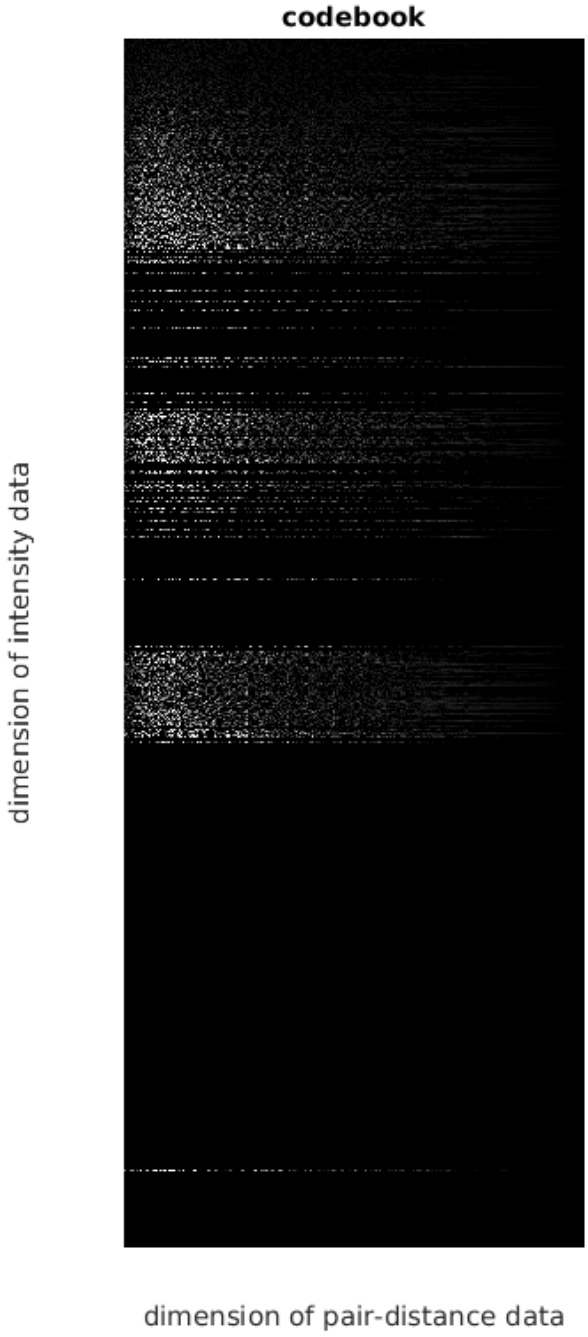

Fig. 3. The estimated codebook - transformation matrix for scattering intensities and pair-distance distributions.

Acknowledgements

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.

References

Lynd, L. R. et al. (1991). *Science* **251**, 1318–1323.

Himmel, M. E. et al. (2007). *Science* **315**, 804–807

Himmel, M. E. (2009). *Biomass recalcitrance: deconstructing the plant cell wall for bioenergy.* Wiley-Blackwell.

Mok, W. S. L., Antal Jr, M. J. (1992). *Industrial Engineering Chemistry Research* **31**, 1157-1161

Nishiyama, Y. et al. (2002). *Journal of the American Chemical Society* **124**, 9074-9082

Nishiyama, Y. et al. (2003). *Journal of the American Chemical Society* **125**, 14300-14306

Driemeier, C. et al. (2011). *Cellulose* **18**, 1509-1519

Pingali, S. V. et al. (2010). *Biomacromolecules* **11**, 2329-2335

Fernandes, A. N. et al. (2011) *Proc. Natl. Acad. Sci.* **108**, E1195-E1203

Thomas, L. H. et al. (2013) *Plant Physiol.* **161**, 465-476

Inouye, H. et al. (2014) *Scientific Reports* **4**

Zhang, Y. et al. (2015) *Cellulose* **22**, 1495-1504

https://debyer.readthedocs.io/en/latest/usage

Nishiyama, Y. et al. (2012). *Cellulose* **19**, 319-336

Inouye, H. et al. (1993). *Biophysical Journal* **64**, 502-519

Beckham, Gregg. et al. (2011) *Current Opinion in Biotechnology* **2**, 231-238

Beckham, Gregg. et al. (2011) *The Journal of Physical Chemistry B* **115**, 4118-4127

Rabideau, B. D. et at. (2013) *The Journal of Physical Chemistry B* **117**, 3469-3479

Bu, L. et al. (2015) *Carbohydrate Polymers* **125**, 146-152

Matthews, J. et al. (2015) *Journal of Chemical Theory and Computation* **8**, 735-748

Zhang, Y. et al. (2016) *Journal of Applied Crystallography* **49**, 2224-2248.

Brooks, B.R. et al. (2009) *J. Comput. Chem.* **30**, 1545-1614

Hynnien, A.P. et al. (2014) *J. Comput. Chem.* **35**, 406-413

Langan, P. et al. (2005) *Cellulose* **12**, 551-562

Bracewell, R. N. (1986) *The Fourier transform and its applications* New York: McGraw-Hill

Synopsis

A codebook is trained and generated as the inverse Fourier Transform of scattering intensities. A reconstructed pair-distance distribution is the multiplication of input intensity and codebook.