3-D statistical characterization of the heterogeneity of biological macromolecular complexes by electron microscopy

Nan Xu[1], **Yunye Gong**[1], **Yili Zheng**[2], **Qiu Wang**[1,3], **Peter C. Doerschuk**[1]

[1] *Electrical and Computer Engineering, Phillips Hall, Cornell University, Ithaca, NY 14850 USA*
[2] *Lawrence Berkeley National Laboratory, 1 Cyclotron Road Mail Stop 50A1148, Berkeley, CA 94720-8142 USA*
[3] *now at: Siemens Corporate Research, 755 College Road E, Princeton, NJ 08540 USA*
[nx25,yg326, qw32, pd83]@cornell.edu, yzheng@lbl.gov

Abstract: Electron microscopy provides images of macromolecular complexes from which the 3-D structure of the complex can be computed when all instances of a complex are identical. An algorithm for characterizing the 3-D spatial statistics of the complex is described and demonstrated for the important case when different instances are not identical.

OCIS codes: 100.3010, 100.3190, 110.0180, 110.1758

1. Introduction

Biology includes many macromolecular complexes in the size range of roughly 10^1–10^2 nanometers, such as viruses, chaperonins, ribosomes, apoptosomes, transcription factor IID, *etc*. These nanoscale machines typically self assemble and have a variety of "stages" in their existence. They sometimes have variable stoichiometry of their chemical constituents. They typically are subject to motions and these motions are central to the functioning of the machine, such as a ribosome moving down a mRNA molecule as the triplets of mRNA bases are translated into amino acids in the growing peptide polymer.

Electron microscopy, especially in the form of single-particle cryo electron microscopy (cryo EM), is transforming structural biology, i.e., the study of the geometry of the macromolecular constituents and how the geometry influences the function of the machine. In these experiments, individual complexes (often described as "particles") are imaged, so the ability to create a crystal (2-D or 3-D) is not required. Usually the images are recorded from unstained particles in order to minimize distortion so contrast is low. The image is roughly a tomographic projection of the particle's electron scattering intensity along the optical axis of the microscope. The fact that the electron beam rapidly damages the particle has several implications. (1) The beam current is minimized leading to images with SNR substantially less than one. (2) The experiment is done at cryogenic temperatures in order to minimize the damage and the freezing is done at a fast rate in order to create vitreous solid water rather than ice since the increase in volume from liquid water to ice would tear apart the particles. (3) Only one image is taken per particle so a 3-D reconstruction requires combining information from many particles. In many experiments it is not possible to have any control over the orientation of the particle as it is placed in the microscope. Furthermore, the poor SNR makes it difficult to determine the orientation from the image. Therefore, reconstruction generally involves both determination of the projection orientation of each image as well as the 3-D structure of the particle.

The standard assumption is that the particles are essentially identical, or at worst come from a few classes of particle and each particle in a class is essentially identical, and we describe this situation as a homogeneous ensemble. However, when different particles have variable stoichiometry or when motion is important, then there are differences between particles even within a single particle class, and we describe this situation as a heterogeneous ensemble and characterizing the heterogeneity is important for understanding the functioning of the particle. A statistical point of view for characterizing heterogeneity has been developed [1] which has many attractive properties relative to the primary alternative which is a resampling method [2] and current work is described.

2. Statistical image processing

For a wide range of resolution, the ideal image is a linear function of the 3-D electron scattering intensity of the particle which is described as a linear combination of known basis functions. Therefore, if z_i is the ith noise-free image arrayed as a vector and the weights in the linear combination for the ith image are the vector c_i then there is a matrix $L(\theta_i)$

where θ_i describes the projection orientation (e.g., Euler angles) such that $z_i = L(\theta_i)c_i$. The first source of uncertainty in the image is θ_i. The second source of uncertainty is pixel noise which includes electron counting noise (the typical dose is tens of electrons per pixel) but also includes many other causes such as variability in the ice. In order to arrive at a computable solution, we have used an additive zero-mean white Gaussian noise model where the noise variance is Q. The third source of uncertainty (our primary innovation) is the description of particle heterogeneity by describing c_i as an i.i.d. realization of a Gaussian random vector with unknown mean vector $\bar{c}^{\eta_i}$ and unknown covariance matrix V^{η_i} where η_i is a discrete random variable that is the class of the ith image. The fourth source of uncertainty is η_i. Let y_i be the experimental image arrayed as a vector. Then $y_i = z_i + w_i = L(\theta_i)c_i + w_i$ where w_i is the ith realization of the pixel noise and the goal of reconstruction calculations is to determine $\bar{c}$ and V from a collection of y_i ($i \in \{1,\ldots,N_v\}$).

If $y_i = c_i$ and the goal is to determine $\bar{c}^{\eta}$ and V^{η} from a collection of y_i by maximum likelihood (ML) estimation then there is a standard expectation-maximization algorithm [3] which uses η_i as the so-called nuisance parameters. Our work is a generalization of this algorithm. In particular, the nuisance parameter is now (θ_i, η_i), there is a linear transformation L, and there is an additive noise w_i. One point of view on this generalization is that we are attempting to partition the variability in the images among multiple sources: orientation (θ_i), class (η_i), pixel noise (w_i), and heterogeneity (c_i). In this point of view, one would expect that it is necessary to estimate the pdf on θ_i, q_j (the pmf on η_i), Q (the variance of w_i), and $\bar{c}^{\eta}$ and V^{η} (the mean and covariance of c_i) even though only $\bar{c}^{\eta}$ and V^{η} are of biological interest. While we have an algorithm capable of doing all of this, in our calculations to date it has not been necessary to estimate the pdf on θ_i, possibly because the virus particles we have worked with are approximately spherical in shape so that assuming a uniform pdf on θ_i, i.e., Haar measure on SO_3, is an sufficiently accurate approximation.

3. Statistics with symmetry, understanding the reconstruction, and parallel computing

Some particles (many viruses) have symmetry. For a particular class, let $\rho(\mathbf{x})$ be the electron scattering intensity as a function of 3-D coordinates $\mathbf{x}$ in real space and R_β ($\beta \in \{1,\ldots,N_g\}$) be the set of 3×3 orthonormal rotation matrices that describe the symmetry. In a homogeneous ensemble, symmetry of ρ means $\rho(\mathbf{x}) = \rho(R_\beta^{-1}\mathbf{x})$ for $\beta \in \{1,\ldots,N_g\}$. This requirement can be built into the basis functions by choosing functions that transform as the identity representation of the symmetry group. In a heterogeneous ensemble, ρ is random with mean $\bar{\rho}(\mathbf{x})$ and correlation function $r_\rho(\mathbf{x}_1,\mathbf{x}_2)$ and symmetry of ρ means $\bar{\rho}(\mathbf{x}) = \bar{\rho}(R_\beta^{-1}\mathbf{x})$ and $r_\rho(\mathbf{x}_1,\mathbf{x}_2) = r_\rho(R_\beta^{-1}\mathbf{x}_1, R_\beta^{-1}\mathbf{x}_2)$ both for $\beta \in \{1,\ldots,N_g\}$. This is a more complicated situation and requires basis functions that transform as all irreducible representations of the symmetry group and requires a structure on V. However, if the biology questions can be answered by the space-varying variance of ρ alone, i.e., $s(\mathbf{x}) = r_\rho(\mathbf{x},\mathbf{x})$, then the situation is simpler. In particular, $s(\mathbf{x}) = s(R_\beta^{-1}\mathbf{x})$ for $\beta \in \{1,\ldots,N_g\}$ so using only basis functions that transform as the identity representation of the symmetry group will achieve the necessary symmetry.

The correlation function $r_\rho(\mathbf{x}_1,\mathbf{x}_2)$ is a large amount of information because it depends on $2\times 3 = 6$ independent variables. Rather than visualize r_ρ, we are attempting to understand r_ρ by computing mathematical mechanical models such that the equilibrium statistical mechanics correlation function of the model matches the correlation function estimated from the image data and then computing properties (such as normal modes) of the mechanical model.

High spatial resolution is always a goal in structural biology, at least to the order of 2Å resolution. Such goals, not yet achieved since current high resolution homogeneous reconstructions are at approximately 4Å resolution, require large numbers of images and basis functions and therefore substantial computation. The algorithm's primary calculation is integration over the nuisance parameters. There are at least two opportunities for parallelism: partition the region of integration and integrate over subregions in parallel (which is in our current software) and partition the set of images and compute in parallel on different subsets. Some of the calculations, such as evaluation of basis functions, appear to be ideally suited for general purpose GPUs. Current software development focuses on incorporation of parallelism at the multicore, GPU, and multi compute node levels via C, C++, OpenMP, CUDA, and MPI in an effort to take advantage of all opportunities that may be available in a user's computing system.

References

1. Y. Zheng, Q. Wang, and P. C. Doerschuk, "3-D reconstruction of the statistics of heterogeneous objects from a collection of one projection image of each object," J. Opt. Soc. Am. A **29**, 959–970 (2012).
2. W. Zhang, M. Kimmel, C. M. T. Spahn, and P. A. Penczek, "Heterogenneity of large macromolecular complexes revealed by 3D cryo-EM variance analysis," Structure **16**, 1770–1776 (2008).
3. R. A. Redner and H. F. Walker, "Mixture densities, maximum likelihood and the EM algorithm," SIAM Review **26**, 195–239 (1984).