Structure from Random Snapshots

Abbas Ourmazd\textsuperscript{1}, Peter Schwander\textsuperscript{1}, and George N. Phillips, Jr.\textsuperscript{2}
\textsuperscript{1}Dept. of Physics, University of Wisconsin, Milwaukee, WI 53211, USA
\textsuperscript{2}Dept. of Biochemistry and the Center for Eukaryotic Structural Genomics, University of Wisconsin, Madison, WI 53706, USA

ABSTRACT

Advanced manifold-based techniques can be used to determine the structure and dynamics of an evolving object from a collection of ultra-low-signal, two-dimensional snapshots emanating from unknown orientations and conformations.

Keywords: structure determination, electron microscopy, X-ray scattering, biomolecules

1. INTRODUCTION

At least in biology, it is well-known that structure determines function, and there is increasing evidence that structure is neither immutable, nor static. Structural variability\textsuperscript{1} and dynamics\textsuperscript{2} have now been observed in ribosome particles and many other well-characterized systems.\textsuperscript{3} Brink et al. highlighted the conformational variations of human fatty acid synthase, an important enzyme in metabolism.\textsuperscript{4} Yu et al. showed that reversible, pH-driven conformational changes of flaviviruses are central to the mechanism by which they are processed in the infection of host cells.\textsuperscript{5} A high-resolution study of the GroEL particles that help other proteins achieve their function configurations revealed significant deviations from existing crystal structure analyses.\textsuperscript{6} Even in crystals, a new treatment of data has revealed statistically significant evidence for the presence of an ensemble of conformations.\textsuperscript{7} The study of structural variability and dynamics thus represents an important, but difficult frontier in understanding biological processes at the atomic level.

In principle, "single-particle" approaches are ideally suited to study structural variability. Cryo-EM has been the single-particle technique of choice, yielding a host of valuable information.\textsuperscript{8,9} The ultimate quality of cryo-EM data, and hence achievable resolution are determined by radiation damage, noise and conformational variability in the particle ensemble. Emerging XFEL methods have used intense short pulses containing up to $10^{12}$ photons to obtain coherent diffraction snapshots of individual particles before the particle is destroyed.\textsuperscript{10} By collecting data before significant damage has occurred, these so-called "scatter-and-destroy" approaches\textsuperscript{11-16} promise to mitigate radiation damage. Several experimental capabilities are needed for this promise to be realized. These include maintaining the native state of biomolecular assemblies injected into the X-ray beam, collecting data with sufficient signal-to-noise ratio before significant radiation damage, and the availability of robust algorithms for reconstructing the three-dimensional (3D) structure from noisy 2D snapshots in the presence of background scattering and intrinsic structural variability in the particle ensemble. Recent experimental results obtained at the FLASH soft-X-ray FEL in Hamburg, Germany\textsuperscript{16-19} have shown that diffraction snapshots of single biological particles can be obtained in their native state, demonstrating that the necessary experimental techniques are within reach. Simulations indicate the effect of radiation damage to be below the 0.3nm level.\textsuperscript{20} The on-going extension of this capability to the hard X-ray regime with the Linear Collider Light Source (LCLS) at the Stanford Linear Accelerator Center (SLAC) offers an unprecedented opportunity to determine the 3D structure of macromolecular assemblies to high resolution.

Radiation damage (or particle destruction) and weak scattering severely limit the signal in each snapshot. Cryo-EM and XFEL approaches can require $\sim 10^9$ snapshots from identical copies of a biological object to reconstruct its structure. In both cases, the relative orientations of the snapshots are unknown. Even in the absence of structural variability, determining the orientations is a key challenge, because the signal-to-noise ratio in each snapshot is so poor.\textsuperscript{21,22} The investigation of structural variability, be it due to different conformations or ligand binding states in the ensemble of particles, complicates matters and represents a further serious challenge to single-particle approaches.\textsuperscript{1,21} The study of
structural variability by cryo-EM has relied on data from other techniques and/or ad hoc assumptions. Supervised classification, the most commonly used approach, sorts the snapshots according to similarity to reference templates, and thus requires prior knowledge of the number and type of structural classes present. 23,24 A recent statistically principled but computationally expensive Expectation-Maximization-based study of structural variability had to resort to trial-and-error to estimate the number of conformations. 1 Nonetheless, this study highlighted the power of algorithmic approaches, which naturally treat structural and orientational variability on an equal footing, and exploit the information content of the entire dataset at each step. The possibility of determining the 3D structure of conformationally heterogeneous objects by XFEL methods is new. 25,26 The combination of paucity of signal, lack of orientational information, and structural variability constitutes a formidable challenge.

2. STRUCTURE RECOVERY IN THE PRESENCE OF NOISE

In essence, the power - and resolution limit - of any reconstruction approach is determined by its ability to extract information from the noisy dataset. A measure of the efficiency of different approaches is the lowest signal at which each can determine snapshot orientations in the absence of structural variability. Cryo-EM snapshots can be oriented down to a mean electron count of ~10 /Å^2, with 30 representing a typical value. 6 The presence of symmetry is often exploited, increasing the effective electron count by the number of symmetry elements, which can be as 20 or more for the high symmetry, icosahedral particles. Shneerson et al showed that orienting XFEL coherent diffraction snapshots by the "common-line" approach 27 requires ~1000x more signal than is available. 22 The first demonstration of structure recovery from simulated XFEL snapshots of macromolecules was published in January 2009 (online publication; Nov. 2008). 26 In Sept. 2009, Loh and Elser published another demonstration 28 using a fundamentally similar approach. 29 Key to success in both these studies is the realization that the information content of the entire dataset must be used at each step. This is because each snapshot contains information about every other, much as the picture from the back of a person's head provides information about the position of the ears, and thus contributes to reconstructing a full-frontal image. This approach was used to reconstruct the 3D structure of a macromolecule from simulated XFEL snapshots of unknown orientation. 26 Using a somewhat simplified model, Elser has argued that the type of approach we have used is capable of operating at even lower signal levels. 30 It is clear that approaches exploiting the information content of the entire dataset extract signal from noise with extreme efficiency, pointing the way to 3D structure recovery to unprecedented resolution with established and emerging single-particle techniques.

Figure 1. The manifold expresses the information content of the dataset. A molecule has only three orientational degrees of freedom. This means that the p pixel intensities in a snapshot change in a correlated fashion with molecular orientation. This correlation is described by a 3D manifold in the p-dimensional space of pixel intensities.
3. MANIFOLD EMBEDDING

Of the approaches able to extract information from the entire dataset at each step of an algorithm, those based on the concept of manifolds are particularly powerful. To appreciate the concept, consider an object able to assume any orientation in 3D space, with each snapshot stemming from an unknown orientation of the object. A snapshot consisting of $p$ pixels can be represented as a $p$-dimensional vector, with each component representing the intensity value at a pixel (Fig. 1). The fact that the intensities are a function of only three orientational parameters ("Euler angles") means that the $p$-dimensional vector tips all lie on a 3D manifold in the $p$-dimensional space of intensities. This manifold, which represents the information content of the dataset, is traced out by the correlated way in which the $p$ pixel intensities change with particle orientation. Each point on the manifold represents a snapshot at a particular orientation. Determining this manifold allows one to assign an orientation to each snapshot. A number of powerful techniques have been developed to discover low-dimensional manifolds in high-dimensional data. Each has its strengths and limitations, with the most common problem being noise sensitivity. We have developed noise-robust versions of Generative Topographic Mapping (GTM), Isomap, and Diffusion Map, demonstrating structure recovery at $\sim 10^{-2}$ photons/detector-pixel. GTM is computationally the most expensive, but has the advantage of allowing one to specify the key variables of the problem (in this case the "Euler angles") as dimensions of a so-called "latent" space, which is then embedded in the "manifest" space of the data. The achievable resolution for structure recovery depends on noise, type of algorithm, and available computational resources. Using simulated snapshots from a set of identical objects in unknown orientations and assuming GTM running on a 100-node cluster of 2.33GHz Intel Core 2 Duo processors, we have shown that it should be possible to recover the structure of a 500kD molecule to 0.3nm, a 1MD molecule to 0.4nm, and a 2MD molecule to 0.5nm. Recent, as-yet unpublished conceptual advances have extended the object size, achievable resolution, and acceptable signal-to-low ratios further.

Manifold embedding techniques exploit the information content of the entire dataset at every step. GTM has the additional capability to generate a snapshot corresponding to any specified point on the manifold - hence the "generative" designation. The reconstructed snapshot, whether a diffraction pattern or image, is not simply an average over the snapshots assigned to an orientational bin as in standard classification approaches, but stems from the entire data set. This produces signal extraction capabilities superior to approaches which classify images and then form class averages to reduce noise.

4. STRUCTURE AND CONFORMATION FROM 2D SNAPSHOTS

We now show that when the snapshots emanate from different discrete conformations, manifold embedding automatically sorts them into separate classes and determines their orientations. This approach, also applicable to cryo-EM images, allows one to reconstruct the 3D structure of each conformation - and, by extension, species - separately. Fig. 2 shows the results when a mixture of randomly oriented diffraction snapshots from the closed and open conformations of the molecule adenylate kinase (ADK, Protein Data Bank designations: 1ank and 4ake, respectively) are presented to noise-robust manifold embedding algorithms at a signal level corresponding to 4x10^{-2} photons/pixel at 0.18nm. Because of their chemical identity, the conformations of ADK are extremely difficult to separate chemically.
Figure 2. Manifold embedding separates snapshots from different conformations and finds the orientations within each set with no a priori knowledge. When a mixture of diffraction snapshots from the molecule ADK in its open and closed conformations is presented to noise-robust versions of GTM or Isomap at signal levels corresponding to 0.04 ph/pixel@0.18nm, the snapshots are automatically sorted into different manifolds and their orientations determined. The 8.5-σ separation between the two manifolds implies extreme fidelity in separating different conformations.

As shown in Fig. 2, manifold embedding automatically sorts the snapshots into separate manifolds, and determines their orientations to within a Shannon angle. We note that no prior information was provided to the algorithm regarding the type or number of conformations. The confidence level with which sorting was performed can be deduced as follows. Noise causes the vectors representing the snapshots to depart from the noise-free manifolds, thus giving a certain "thickness" to each manifold. The sorting confidence can be deduced from the closest separation between the two manifolds expressed in standard deviations of the distributions of vectors about the manifolds. At the signal level of 4x10^{-2} photon/pixel with Poisson noise, the smallest separation between the two manifolds exceeds 8.5 standard deviations. This means that snapshots from the different conformations are sorted with extreme fidelity. We note that larger objects such as macromolecular assemblies produce larger signals. It should therefore be possible to use manifold embedding to map their conformations with even greater precision. Manifold embedding can thus sort discrete molecular conformations - and, by extension, different species - with extreme fidelity, offering a convenient route to "post facto" purification of solutions used in single-particle experiments.

We next consider the unfolding of a molecule to demonstrate the principle of mapping conformational continua. The unfolding of ADK was simulated by molecular dynamics as follows. The coordinates of ADK from E. coli in the open state (Protein Data Bank designation: 1ake) were placed in a spherical droplet of water and simulated at a nominal temperature of 850 K using NAMD.\textsuperscript{46} 12,500 diffraction snapshots were simulated from 100 conformations, with each conformation assuming 125 orientations about one axis. Snapshots were provided to a modified version of the Isomap manifold embedding algorithm, and the resulting manifold displayed through its projections along the first three dominant eigenvectors (Fig. 3).
It is clear that orientational and conformational variations give rise to a tubular manifold. Qualitatively, the closed cross-sections of the tube include orientational change, while paths terminating at the tube ends indicate conformational change. In order to separate an orientational change from a conformational change, however, the directions corresponding to pure orientational and pure conformational change must be identified at each point on the manifold. This can be achieved by recognizing that the manifold is Riemannian. Due to the SO(3) symmetry of molecular orientations, the Killing vectors on the manifold point in directions of pure orientational change, and thus also identify the directions of pure conformational change. Manifolds with SO(3) symmetry in some directions have received considerable attention in general relativity and lattice space-time, and well-established techniques exist for determining their Killing vector fields. Noise-robust versions of Riemannian techniques can thus be used to identify directions of pure orientational and conformational change on the manifold, and thus reconstruct the 3D structure of the conformations of macromolecular assemblies.

5. CONCLUSIONS

New approaches to structure recovery represent the information content of a dataset as a Riemannian manifold, whose properties can be determined by techniques developed in differential geometry, general relativity, and graph theory. In this picture, the structure and conformations of an object are known, if, given any 2D snapshot, any other can be recovered. This requires the ability to “navigate” toward specific points on the manifold and produce high-quality snapshots corresponding to each point. Understanding “perception” in terms of the ability to navigate a manifold offers a powerful route to studying the structure and dynamics of objects on unprecedented length and timescales.
ACKNOWLEDGEMENTS

This work was performed in collaboration with Dimitrios Giannakis (New York University). We are grateful to Mary Eckert for help with preparation of the manuscript.

REFERENCES