Mapping the conformations of biological assemblies

This article has been downloaded from IOPscience. Please scroll down to see the full text article.

2010 New J. Phys. 12 035007

(http://iopscience.iop.org/1367-2630/12/3/035007)

The Table of Contents and more related content is available

Download details:
IP Address: 129.89.57.227
The article was downloaded on 01/04/2010 at 15:43

Please note that terms and conditions apply.
Mapping the conformations of biological assemblies

P Schwander\(^1\), R Fung\(^1,3\), G N Phillips Jr\(^2\) and A Ourmazd\(^1,4\)

\(^1\) Department of Physics, University of Wisconsin Milwaukee, 1900 E Kenwood Blvd, Milwaukee, WI 53211, USA
\(^2\) Department of Biochemistry and the Center for Eukaryotic Structural Genomics, University of Wisconsin Madison, 433 Babcock Drive, Madison, WI 53706, USA
E-mail: ourmazd@uwm.edu

Received 28 September 2009
Published 31 March 2010
Online at http://www.njp.org/
doi:10.1088/1367-2630/12/3/035007

Abstract. Mapping conformational heterogeneity of macromolecules presents a formidable challenge to x-ray crystallography and cryo-electron microscopy, which often presume its absence. This has severely limited our knowledge of the conformations assumed by biological systems and their role in biological function, even though they are known to be important. We propose a new approach to determining to high resolution the three-dimensional conformations of biological entities such as molecules, macromolecular assemblies and ultimately cells with existing and emerging experimental techniques. This approach may also enable one to circumvent current limits due to radiation damage and solution purification.

\(^3\) Present address: Department of Physics and Astronomy, University of California Los Angeles, 475 Portola Plaza, m/c 154705, Los Angeles, CA 90095, USA.
\(^4\) Author to whom correspondence should be addressed.
1. Introduction

The biological functions of macromolecules and their assemblies almost always involve conformational changes, which can be quite pronounced or very subtle. High-resolution determination of macromolecular conformations is thus an important scientific frontier. In principle, single-particle approaches are ideally suited to determining conformations of biological entities such as molecules, molecular assemblies, chromosomes and cells. In reality, however, single-particle techniques often rely on ‘averaging’ data obtained from a large ensemble of objects usually assumed to be identical. Conformational information on biological systems thus remains incomplete. Nonetheless, there is increasing appreciation that the dynamic behavior of macromolecules is an inherent part of their functional design. Examples range from the classical descriptions of the R to T states in hemoglobin and related atomic motions in myoglobin [1] to the recent recognition that the virulence of the dengue virus strongly depends on transitions in its protein contacts and conformational rearrangements [2].

Despite powerful contributions to the study of proteins and some assemblies, x-ray crystallography and nuclear magnetic resonance (NMR) have limitations. With notable exceptions, the constraints imposed by crystals have limited the role of x-ray crystallography in elucidating conformational variety. NMR, while able to study conformations in biomolecules of modest size, has not been extensively applied to larger systems. Cryo-electron microscopy (cryo-EM) has been extensively used to study nominally identical macromolecular assemblies [3]–[6]. When conformational variety has been explicitly addressed, the results, won with effort and ingenuity, provide tantalizing evidence of a rich variety of conformations, even in well-studied systems [5]–[7]. A deep understanding of the nature and role of conformational variety in biological function would revolutionize our knowledge of key processes ranging from basic cell function to pathological states. Unraveling the role of conformations in virulence, for example, is expected to lead to new strategies for fighting infection.
Here, we show that a new generation of algorithms combining techniques from Riemannian geometry, general relativity and machine learning is poised to deliver a powerful new approach to biological structure determination in a way that maps conformational heterogeneity, and circumvents longstanding limits due to radiation damage and noise. These algorithms can be used with established experimental techniques, such as cryo-EM, and emerging approaches exploiting the extreme brightness of x-ray free electron lasers (XFELs). By transforming the limits set by the nature of biological entities—weak scattering, radiation damage, conformational variety—to limits set by computational resources, which have improved exponentially for 50 years, these algorithmic approaches promise a decisive advance in biostructure determination.

In the long term, these approaches are expected to constitute a new platform for determining structure and conformations in ways that mitigate the limits set by the nature of biological entities, and by vexing experimental issues such as solution purification and crystallization. The work described in this paper is the next vertical step along a continuum beginning with our demonstration that the structure of individual macromolecules can be determined to high resolution from an ensemble of coherent diffraction snapshots of unknown orientation at mean photon counts as low as $\sim 10^{-2}$ per Shannon pixel \[8\] (figure 1). This is orders of magnitude below the signal levels previously required \[9\]. The algorithms that we have developed have closed a critical gap in proposed techniques for determining the structure of individual biological entities, whereby a succession of identical particles is exposed to single, short and intense pulses from an XFEL source \[10, 11\]. We are now in a position to address the next step: the hitherto unanticipated possibility of determining to high resolution the three-dimensional (3D) structure of the conformations of an object from random snapshots of an ensemble of non-identical objects, each in a different conformation. This is expected to have a substantial impact on existing cryo-EM techniques \[3\] and the proposed XFEL-based ‘scatter-and-destroy’ approaches \[11\], significantly advancing the tomographic study of increasingly complex systems such as chromosomes and perhaps whole cells.
This paper is organized as follows. Section 2 presents a brief overview of recent trends in determining macromolecular structure and conformations. Section 3 outlines a new algorithmic approach to determining structure, with section 4 summarizing our current understanding of its capabilities. Section 5 addresses the application of this approach to determining macromolecular conformations. Section 6 places the various algorithms in context, indicating possible routes to further progress. Section 7 summarizes and concludes the paper.

2. Overview of current knowledge and techniques

There is increasing awareness that structural variability is not only common, but can play a key role in macromolecular function. Revealing structural variability in macromolecules and their assemblies often requires spatial resolutions beyond the reach of current x-ray tomographic techniques [12]. Applying sophisticated algorithmic approaches to cryo-EM data, Scheres et al revealed structural variability in 70S E. coli ribosome particles [5] and other well-characterized systems [13]. Combining normal mode analysis with cryo-EM, Brink et al [14] highlighted the conformational variations of human fatty acid synthase. Yu et al [2] showed that reversible, pH-driven conformational changes of flaviviruses are central to the mechanism by which they are processed and stabilized in the host cell. A high-resolution cryo-EM study of GroEL revealed significant deviations from existing crystal structures [6]. Even in crystals, variation in the structures can be seen. Often, distinct conformations can be discerned within an electron density map, which is, of course, an average over time and space. Different crystal forms of the same molecule also show variations in the underlying structure, demonstrating that crystallization chooses conformations from a larger ensemble [15]–[17]. Recently, a new treatment of data has revealed statistically significant evidence for the presence of an ensemble of conformations [18]. Although the water content of protein crystals is about the same as in living cells, the crystal lattice can also affect the dynamics by virtue of regular contacts [19]. Indeed, sub-nanosecond time-resolved crystallography of conformational changes can be performed [1]. Despite mounting evidence of its importance, the study of structural variability has proved difficult, limiting our ability to relate structure and dynamics to function.

The ultimate quality of experimental structural data and hence achievable spatial resolution is determined by radiation damage and/or noise. Cryo-EM, the single-particle technique of choice, for example, has yielded a host of valuable information, but is severely limited by radiation damage [20]. Emerging XFEL methods have recently used intense short pulses containing up to $10^{12}$ photons to obtain coherent diffraction snapshots of individual particles before the particle is destroyed [21]. By collecting data before significant damage has occurred, these so-called ‘scatter-and-destroy’ approaches [10], [22]–[26] 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 3D structure from noisy 2D snapshots of unknown orientation in the presence of background scattering. Recent experimental results obtained at the FLASH soft-x-ray FEL in Hamburg, Germany [26]–[29] have shown that diffraction snapshots of single biological particles can be obtained in their native state, indicating that the necessary experimental techniques are perhaps within reach. Simulations indicate the effect of radiation damage to be below the 0.3 nm level [30]. The imminent
extension of this capability to the hard x-ray regime offers an unprecedented opportunity to
determine the 3D structure of macromolecular assemblies to high resolution. The world’s first
hard-XFEL, the Linac Coherent Light Source (LCLS) at the Stanford Linear Accelerator Center
(SLAC), produced the first photons in April 2009, and commenced lasing 10 days later. There
are thus at least two single-particle approaches in principle able to study structural variability at
high resolution. They are both affected by radiation damage.

The combination of radiation damage, weak scattering, unknown snapshot orientations
and structural variability constitutes a formidable challenge. Cryo-EM and XFEL approaches
can require \(\sim 10^6\) 2D snapshots from identical copies of a biological object to reconstruct
its 3D structure. 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 [3, 9]. The
investigation of structural variability, be it due to different conformations or ligand binding
states in the ensemble of particles, further complicates matters [3, 5]. 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 method, sorts the snapshots according to
similarity to reference templates, and thus requires prior knowledge of the number and types
of structural classes present [31, 32]. 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 [5]. 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. It has been
proposed to use as-yet unavailable experimental capabilities, such as simultaneous recording of
multiple projections with femtosecond accuracy, and/or undemonstrated algorithms to recover
the 3D structure of non-identical objects to limited resolution [33, 34]. The possibility of
determining the 3D structure of conformationally heterogeneous objects to high resolution by
XFEL methods is new [8].

The power—and resolution limit—of any reconstruction approach is determined by its
ability to extract information from the noisy dataset. Determining the snapshot orientations
is perhaps the most critical step in 3D structure recovery, because it has to be performed at
extremely low ‘raw-signal’ levels. Different techniques must therefore be judged by the lowest
signal at which they can determine snapshot orientations, in the first instance in the absence
of structural variability. Cryo-EM snapshots can be oriented down to a mean electron count of
\(\sim 10^{-2}\), 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
high as 120 for icosahedral particles viewed in diffraction. Shneerson et al showed that orienting
XFEL coherent diffraction snapshots by the ‘common-line’ approach [35] requires \(\sim 1000\times\)
more signal than available [9]. The group of the corresponding author published the first
demonstration of structure recovery from simulated XFEL snapshots of macromolecules [8].
A second demonstration was recently published by Loh and Elser [36], who recovered
orientations through an expectation–maximization procedure under the constraint of manifold
contiguity implemented as a so-called expansion–contraction cycle. These constitute the
essential features of generative topographic mapping (GTM), the approach used by us [8].
As in [13], key to success 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
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.

was used to reconstruct the 3D structure of a small macromolecule from simulated XFEL snapshots of unknown orientation [8] (figure 1). Using a simplified model, Elser has argued that the type of approach we have used is capable of operating at even lower signal levels [37]. 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.

3. New approach: manifold mapping

Of the several approaches for extracting information from the entire dataset, 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 (figure 2). 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 [8]. A number of powerful techniques has been developed to discover low-dimensional manifolds in high-dimensional data [38]–[41]. Each has its strengths and limitations, with the most common problem being noise sensitivity [42, 43]. Some manifold mapping techniques attempt to determine the underlying dimensionality of the manifold from the data, but suffer from the disadvantage that the manifold topology and dimensionality can be strongly affected by noise [42]. We have developed noise-robust versions of GTM, Isomap and Diffusion Map, demonstrating structure recovery with GTM, Diffusion Map and Isomap at $\sim 10^{-2}$ photon Shannon-pixel$^{-1}$ [44]. 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 2.33 GHz Intel Core 2 Duo processors, we have shown that it should be possible to recover the structure of a 500 kDa molecule to 0.3 nm, a 1 MDa molecule to 0.4 nm, and a 2 MDa molecule to 0.5 nm [8]. We expect algorithmic improvements to extend the object size and achievable resolution further (see section 5.3 below). In summary, using noise-robust manifold mapping techniques, the structure of an object can be determined to high resolution from ultra-low-signal snapshots of identical members of an ensemble, each in an unknown orientation. The case of non-identical objects is discussed in section 5 below.

4. Capabilities of manifold mapping techniques

4.1. Dealing with experimental data

Using simulated data, we have established that our approach is robust against the addition of the background measured in single-particle experiments at the FLASH FEL facility in Hamburg, Germany. As shown in figure 3, the R-factor of the reconstructed diffraction volume degrades with increasing background noise, but this can be reversed by increasing the number of snapshots provided to the algorithm. Using the Advanced Photon Source and a nanofoam as an analogue for a complex biological object, we have established that our approach can orient

\[ R = \frac{\sum_{i} (|F_{i}^{\text{true}}| - c|F_{i}^{\text{deduced}}|)^{2}}{\sum_{i} |F_{i}^{\text{true}}|^{2}}, \]

where \( F_{i}^{\text{true}} \) are the Fourier moduli calculated directly at regularly spaced points on a Cartesian grid, \( F_{i}^{\text{deduced}} \) the corresponding moduli deduced by interpolation from the oriented 2D snapshots used to form the 3D diffraction volume, and \( c \) is a scaling factor obtained by minimizing \( R \) with respect to \( c \).

\[ Figure \ 3. \ Adding \ background \ scattering \ increases \ the \ R-factor, \ but \ this \ can \ be \ reversed \ by \ providing \ more \ snapshots \ to \ the \ algorithm. \ (a) \ Alignment \ accuracy \ degrades \ with \ increasing \ background \ scattering. \ (b) \ R-factor \ between \ the \ correct \ diffraction \ volume \ and \ that \ reconstructed \ in \ the \ presence \ of \ background \ (red \ curve) \ tends \ toward \ that \ without \ background \ (blue \ curve) \ as \ the \ number \ of \ snapshots \ is \ increased. \ Signal: \ 0.04 \ photon \ pixel^{-1}@0.18 \ nm \ with \ Poisson \ noise; \ background \ of \ 0.25 \ photon \ pixel^{-1}: \sim1/3 \ of \ that \ measured \ at \ FLASH \ UV \ XFEL. \]

experimental snapshots to within a Shannon angle [8] down to the lowest scattered photon intensity measured to date (0.08 photon pixel$^{-1}$) [45]. Experiments at lower fluxes and similar studies on cryo-EM data are underway. These results further support the notion that manifold mapping algorithms can deal with experimental data.

4.2. Superior signal extraction

Most manifold mapping approaches exploit the information content of the entire dataset to determine the manifold. GTM has the additional capability of generating a snapshot corresponding to any specified point on the manifold—hence the ‘generative’ designation. The reconstructed image is not simply an average over the snapshots assigned to an orientational bin as in standard classification approaches [46, 47], but stems from the entire dataset. This produces signal extraction capabilities superior to approaches that classify images and then form class averages to reduce noise. To illustrate this and make contact with cryo-EM, we use real-space, rather than diffraction snapshots. GTM was used to sort into 60 orientational bins 3600 simulated electron microscope images of the small protein chignolin in random orientations (figure 4). The snapshots were simulated at a mean electron count of $10^2$ Å$^{-2}$ at a point-to-point resolution of 0.37 nm (Phillips CM20T TEM; 100 kV; Scherzer defocus) (figure 4(b)). Like other orientation recovery techniques, GTM was able to accurately classify the snapshots into orientational bins. At this point, standard approaches average the snapshots within a bin to reduce noise. This produces the images shown in figure 4(c). The images generated by GTM from the manifold are shown in figure 4(d). The root-mean-square difference from noise-free images improves from 9.7 for the raw images in (b) to 1.3 for the averaged images in (c) to 0.65 for the GTM-generated images in (d). This demonstrates the superior signal extraction capability of manifold mapping techniques.

It is not straightforward to relate the mean electron dose used in figure 4 to those typically encountered in cryo-EM. Using the ratio \{image variance/(mean electron dose)$^2$\} as a measure of contrast, we compared the simulated images of chignolin at $10^2$ electrons Å$^{-2}$ (figure 4) with those of GroEL at $32$ electrons Å$^{-2}$, the dose used in a recent cryo-EM study [6]. The differences in the contrasts and symmetries of the two particles indicate that the dose of figure 4 is roughly an order of magnitude lower than that commonly used in cryo-EM. This estimate requires validation with experimental data, because effects such as absorption contrast and substrate noise cannot be easily simulated. Nevertheless, the enhanced performance of manifold mapping over conventional image classification and averaging is enticing. Generation of manifold-based images with other manifold mapping techniques [39, 41] is in principle possible, but has not been demonstrated. The ability to use the information content of the entire dataset for each reconstructed image is of decisive importance for recovering orientation, structure and conformations, in both cryo-EM and XFEL.

5. Mapping conformations

5.1. Discrete conformations and ‘post facto purification’

We now show that when the bioparticle beam in an XFEL experiment contains different discrete conformations, manifold mapping automatically sorts the diffraction snapshots into separate classes and determines their orientations. This approach, also applicable to cryo-EM images, allows one to reconstruct the 3D structures of different conformations—and, by
Figure 4. Generating images from the manifold gives superior results. Simulated cryo-EM images of the protein chignolin at 0.37 nm resolution in three orientations. (a) Noise-free snapshots. (b) Snapshots at a mean electron count of \(10^2\) electrons Å\(^{-2}\). (c) Images obtained by averaging the snapshots assigned to each orientational bin. (d) Images generated from the GTM manifold. Note the superior signal-to-noise ratio when the image is generated from the manifold.

extension, species—separately. Figure 5 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 entries: 1ANK and 4AKE, respectively) are presented to noise-robust manifold mapping algorithms at a signal level corresponding to \(4 \times 10^{-2}\) photons pixel\(^{-1}\) at 0.18 nm. Because of their chemical identity, the conformations of ADK are extremely difficult to separate chemically. As shown in figure 5, manifold mapping (by GTM or Isomap) 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 types 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 \(4 \times 10^{-2}\) photons pixel\(^{-1}\) at 0.18 nm with Poisson noise, the smallest separation between the two manifolds exceeds 8.5
5.2. Conformational continua

Ultimately, one would like to map the entire continuum of conformations assumed by a biological system. We use 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 entry: 4AKE) were placed in a spherical droplet of water and simulated at a nominal temperature of 850 K for 5 ns using NAMD [48]. 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 mapping algorithm, and the resulting manifold displayed through its projections along the first three dominant eigenvectors obtained by Isomap analysis (figure 6). It is clear that orientational and conformational variations combine to produce a tubular manifold. Qualitatively, the closed cross-sections of the tube include orientational change, while paths terminating at the tube ends indicate 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 [9]. It should therefore be possible to use manifold techniques to map their conformations with even greater precision. Manifold mapping can sort discrete molecular conformations—and, by extension, different species—with extreme fidelity. This offers the possibility of using solutions containing multiple species and perform ‘post facto purification’ by sorting the data after the experiment.

Figure 5. Manifold mapping 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 (Protein databank (PDB) entry 4AKE) and closed (PDB entry 1ANK) conformations is presented to noise-robust versions of GTM or Isomap at signal levels corresponding to $0.04 \text{ photon pixel}^{-1}\times0.18 \text{ nm}$, the snapshots are automatically sorted into different manifolds and their orientations determined. The results shown here stem from GTM analysis, with Isomap results being very similar. The 8.5-$\sigma$ separation between the two manifolds implies extreme fidelity in separating different conformations.
Figure 6. Manifold traced out by 100 conformations of an unfolding ADK molecule, each able to assume one of 125 orientations about a single axis. The molecular conformations corresponding to different points on the manifold are shown in the insets.

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 [49], and well-established techniques exist for determining their Killing vector fields. Noise-robust versions of Riemannian techniques can 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. A demonstration of this approach is underway.

5.3. Computational issues

The computational demands of current algorithms rise rapidly with the number of resolution elements $r$, typically as $r^n$ with $2 \leq n \leq 3$, where $r \sim (D/d)^3$, and $D$ and $d$ designate the object diameter and resolution, respectively [8, 36]. The discretization of a conformational continuum into $c$ points increases the computational load by the same factor. We have taken the following steps to reduce the computational requirements of GTM. (i) By processing data in small batches, the memory requirements have been reduced to 2 GB of RAM, which is available at each core of our computing cluster. (ii) A modified approach based on a binary search has been developed, which should reduce the exponent $n$ by 1 and the computational load for a typical calculation by $\sim 10^6 \times$. This is currently being implemented. (iii) By ensuring that the modified code can
be straightforwardly parallelized, we expect to reduce the computational time by another 100×. Isomap and Diffusion Map are computationally less expensive than GTM. We have identified significant load reduction measures for these techniques also. These algorithmic modifications help ensure that the increased computational load due to conformational variability can be handled.

5.4. Data collection

In the absence of conformational variability, structure recovery to high resolution can require $\sim 10^6$ snapshots. Ignoring the mutual information between different snapshots (which reduces the number of snapshots needed), mapping a conformational continuum discretized into 100 points requires 100× more data. The LCLS source and relevant detector operate at 120 Hz and are thus able to deliver $10^7$ shots per day. The planned European XFEL is slated to operate at 30 kHz. It is thus realistic to expect that the data needed to map conformational continua can be indeed collected.

6. Discussion

The majority of algorithms used in cryo-EM and those originally proposed for XFEL single-particle work sort similar snapshots into orientational classes, average over members of each class to boost signal, and finally determine orientations. Three years ago, Shneerson et al [9] showed that structure recovery by such algorithms requires 1000× more flux than is theoretically possible with currently envisaged XFELs. Since then it has become clear that single-particle structure recovery, be it from XFEL diffraction snapshots or cryo-EM images, is most efficiently performed with approaches that exploit the information content of the entire dataset [5, 8, 13]. Analysis of simulated data indicates that such approaches can recover single-molecule structure at anticipated XFEL fluxes, and perhaps at doses lower than currently in use in cryo-EM.

These emerging algorithms can be broadly separated into three classes: (i) information theoretic, (ii) manifold embedding and (iii) manifold mapping. Information theoretic approaches such as GTM use Bayesian inference and an expectation–maximization engine with constraints to ensure contiguity of the manifold in data space. This can be a specified latent space of orientations [8, 38] or a less familiar equivalent, such as the ‘expansion–contraction’ cycle used in [36]. In essence, these algorithms are ‘general’ in the sense that they contain no information about the diffraction process. For this reason, they are computationally inefficient [50]. Manifold embedding approaches such as Diffusion Map and Isomap use graph theoretic means to discover the manifold in data space without recourse to a latent space. They incorporate some information about the diffraction process. Although, their kernel-based approach is computationally more efficient, they are noise sensitive and unable to identify the correct manifold topology without prior denoising [42]. Finally, manifold mapping techniques circumvent embedding altogether, using Riemann geometric techniques to ‘navigate on the manifold directly’ without recourse to external coordinate systems. The nature of the diffraction process can be directly incorporated so as to place constraints on the mapping and the properties of the data manifold [51]. As such, they promise an optimal combination of computational efficiency and noise robustness, although this remains to be demonstrated.
Graph theoretic and Riemann geometric approaches are perhaps most appropriate for studying conformational continua, because they require no a priori knowledge of the dimensionality of the data manifold. In contrast, Bayesian expectation–maximization-based approaches either require prior knowledge of the number of conformations present, or must deduce this by trial and error [5]. More generally, Riemann geometric manifold mapping approaches offer a fundamentally new formulation of scattering with potential access to an array of powerful analytical techniques. The implications of this remain to be fully understood. It is nonetheless intriguing that methods developed in cosmology to study the structure of the universe may be used to investigate the building blocks of life.

7. Summary and conclusions

We have outlined powerful new algorithmic approaches based on concepts from information theory, graph theory and Riemannian geometry, and demonstrated their potential for extracting signal from noise, recovering 3D structure with no orientational information, separating discrete conformations and species, and eventually mapping conformational continua. By naturally incorporating conformational heterogeneity, these algorithms promise to substantially increase the range of systems accessible to single-particle techniques such as cryo-EM, and emerging XFEL-based approaches. At the same time, these algorithms offer a route to transforming fundamental limits due to radiation damage and noise to computational issues, because, down to some as-yet-undetermined limit, signal can be traded against the number of snapshots. As computational resources have increased exponentially for five decades, such a development would constitute a fundamental advance in structure recovery.

Acknowledgments

We are grateful to Dimitris Giannakis, Leonard Parker, Dilano Saldin and Valentin Shneerson for valuable discussions, to Ross Harder and Ian McNulty for providing experimental nanofoam data for our analysis and to Henry Chapman for experimental data on background scattering. This work was supported in part by NIH Protein Structure initiative grant GM074901.

References


[38] Bishop C M 1995 Neural Networks for Pattern Recognition (Oxford: Oxford University Press)


[42] Balasubramanian M and Schwartz E L 2002 The isomap algorithm and topological stability Science 295 7a


[44] Giannakis D, Schwander P and Ourmazd A, to be publised


[51] Schwander P, Giannakis D and Ourmazd A, unpublished