Published in final edited form as:

J Time Ser Anal. 2012 September 1; 33(5): 704–717. doi:10.1111/j.1467-9892.2011.00779.x.

Autocovariance Structures for Radial Averages in Small Angle X-Ray Scattering Experiments

F. Jay Breidt,

Department of Statistics, Colorado State University

Andreea Erciulescu, and

Department of Statistics, Iowa State University

Mark van der Woerd

Department of Biochemistry and Molecular Biology, Colorado State University

Abstract

Small-angle X-ray scattering (SAXS) is a technique for obtaining low-resolution structural information about biological macromolecules, by exposing a dilute solution to a high-intensity X-ray beam and capturing the resulting scattering pattern on a two-dimensional detector. The two-dimensional pattern is reduced to a one-dimensional curve through radial averaging; that is, by averaging across annuli on the detector plane. Subsequent analysis of structure relies on these one-dimensional data. This paper reviews the technique of SAXS and investigates autocorrelation structure in the detector plane and in the radial averages. Across a range of experimental conditions and molecular types, spatial autocorrelation in the detector plane is present and is well-described by a stationary kernel convolution model. The corresponding autocorrelation structure for the radial averages is non-stationary. Implications of the autocorrelation structure for inference about macromolecular structure are discussed.

Keywords

Gaussian process; Kernel convolution; Spatial autocorrelation

1 Introduction

1.1 The importance of structural biology

Structural biology is a branch of science concerned with the molecular structure of biological macromolecules, especially proteins and nucleic acids and their complexes. Molecular structures are of great interest to biologists because biological macromolecules are the basis of the functions taking place in cells, providing specific functionality to tissues and ultimately defining how organisms maintain their physiological health. Molecular structure and function are closely related: even minor changes in molecular structure can affect their function.

The number of possible molecules or complexes of interest in structural biology applicable to humans has been estimated to be in the thousands [12]. However, it is not yet known for each molecule or complex what its functions are. The number of molecules which have been studied with the aim to understand and ultimately alter functional characteristics remains

low compared to the number of possible targets. By extension, our knowledge of the structures and functions of biological macromolecules is not yet adequate for application to the study of most diseases, even if in some areas progress has been made.

Numerous examples illustrate the power of structural biology in working towards an answer for an important biological question. For example, consider the compaction of DNA into the cell nucleus. Human cells contain DNA that measures, when fully extended, approximately two meters. This DNA is ultimately contained in a ten-micrometer diameter nucleus. DNA is initially packed into nucleosomes which are then thought to further compact into higher-order (and larger) structures, building chromatin and ultimately chromosomes (compaction and dynamics are reviewed in [18, 8]). However, how exactly the nucleosomes compact, which proteins influence or facilitate this compaction and how they perform this task, i.e. what the exact molecular interactions are, remains unknown. The molecular structure of the nucleosome is already complex: it consists of two copies each of the histones H2A, H2B, H3 and H4, plus approximately 146 base pairs of double stranded DNA that form $1\frac{3}{4}$ turns of DNA around the protein core [19]. Examples of additional proteins that interact with the nucleosomes and have a role in its compaction are histone H1 [29] and MeCP2 [20]. The compaction of DNA into a small particle is only half of the story: DNA also needs to stay accessible to the cellular transcription and replication machinery so the appropriate genes may be activated or copied. This functionality requires that proteins that compact DNA are removed or counteracted by other proteins. The maintenance of compact, yet accessible DNA therefore requires a complex interplay of several factors with the basic building block of the nucleosome. This makes the study of nucleosomes in complex with additional proteins a compelling effort.

1.2 Structural biology and SAXS

Biological macromolecules are too small to see in detail even with the most advanced light microscopes. The methods that structural biologists use to determine their structures generally involve measurements on vast numbers of identical molecules at the same time. The two most common methods used to acquire the structures are X-ray crystallography and nuclear magnetic resonance (NMR) spectroscopy. Each of these methods has its own advantages and disadvantages. One of the (shared) advantages is that these techniques provide high-resolution structures, often with atomic detail. The main disadvantage of NMR is that it has limits to the size and complexity of the molecular species studied, while the main disadvantage to X-ray crystallography is that crystals need to be grown and in the case of macromolecules this can be a formidable task. In practice it is not always feasible to grow these crystals. Small angle X-ray scattering (SAXS) is an alternative method with nearly opposite characteristics: it is experimentally a simple technique, applicable to the majority of molecules, and it provides low resolution information.

In the analysis of biological macromolecules, SAXS is used as a method to analyze dilute solutions. A small amount of the sample (15–30 μ l) of interest, typically in a biological buffer, is exposed to a high-intensity X-ray beam and the scattering pattern is recorded on a two-dimensional detector as shown in Figure 1. The amount of sample needed, the concentration (typically in the range of 2–10 mg/ml) and the exposure time (typically 0.5–60 seconds), allow for a quick and efficient method of data acquisition. A reference experiment is performed that is identical to the main experiment, but different in that the molecule of interest is absent. It provides a background reference that captures the characteristics of the instrument and of the biological buffer. Thus, by an appropriate subtraction of reference experimental data from the data of interest, information about the target molecule can be isolated.

X-rays, just as visible light, can be thought of as a stream of photons. The recording of these photons is described as ‘counting’ in the sense that the intensity of a recorded pixel is assumed to be proportional to the number of accumulated photons. However, in practice a detector has a linear range and observations outside that range should not be incorporated into the data sets. Therefore it is common to perform an experiment in two steps: one short exposure to ensure that the recorded signal is within the detector’s linear range and one longer exposure to optimize weak signals. After the experiments are done, the data from these experiments are combined. A second concern with the use of X-rays is that X-rays are a form of ionizing radiation and are capable of physically altering the samples. Repetitive exposures of the same sample are used to reveal if a sample changes over time (as a function of the acquired X-ray dose). Ideally a sample should not change over time and the repeatedly recorded signals should be the same within a margin of error.

The molecules in the sample are assumed to take on random orientations and positions during the exposure period, and it is assumed that all angular dependence is averaged out uniformly. Subsequently, the two-dimensional images are reduced to one-dimensional scattering curves by radial averaging, i.e. over ϕ in the illustration in Figure 1. The images are then inspected for signs of radiation damage, pixel saturation and the effects of solute concentration. In case of radiation damage, the data are discarded. If saturation occurs, images with different exposure times are merged, eliminating the saturated pixels. When concentration effects are evident, the data are variously ‘extrapolated to zero concentration’, discarded, or used at the lowest concentration measured.

SAXS data interpretation can be divided into two categories: estimation of physical parameters and shape reconstruction. The former involves a handful of scalar quantities and the latter involves a complete three-dimensional envelope for the molecule.

1.2.1 Estimation of physical parameters—Physical parameters are related to molecular size and shape and include radius of gyration, maximum linear dimension of a particle in solution and particle volume. SAXS data interpretation for such parameters is generally applicable to all particles in solution (not just biological molecules) and has a long history, having been conceived and described by Guinier [6].

The radius of gyration is defined via

$$R_g^2 = \frac{\sum m_i r_i^2}{\sum m_i}, \quad (1)$$

where m_i is the mass of the i^{th} atom in the molecule and r_i is the distance of that atom to the center of mass of the molecule. The radius of gyration, together with the scattering intensity at zero angle, describe the beginning of the scattering curve [5] and, *vice versa*, the beginning of the scattering curve can be used to obtain $I(0)$ and R_g . The comparison of R_g values allows us to understand some very basic principles about particle shape, because the R_g value of a sphere and an extended rod-shaped particle (of the same atomic composition) will be significantly different.

The maximum particle dimension is the largest linear dimension within a particle. It is obtained by an indirect inverse Fourier transformation (see example below). Finally, it is possible to derive the volume of a particle in solution directly via Porod analysis [22]. The way and extent that measurement uncertainties affect these parameters directly derived from scattering images is currently not clear.

1.2.2 Shape reconstruction—Shape reconstruction of biological macromolecules was pioneered in the last twenty years and has only been in common practice for approximately a decade (for a comprehensive review of methods, see Putnam *et al* [23]). Eliminating many details, the method can be summarized as follows: the SAXS data are processed by taking (at least) two two-dimensional images (one sample and one reference image), scaling them together guided by the measured intensity of the incident X-ray beam, subtracting the two scaled images, averaging the scattering data radially and condensing the information into a one-dimensional curve (see example in Figure 2). In fact, this one-dimensional curve estimates the squared modulus of the Fourier transform of the electron density of a molecule (e.g., [2], §13.1). This fundamental physical quantity would completely describe the structure of the molecule, but is not identifiable from only the squared modulus. One approach is to calculate a regularized inverse Fourier transform to estimate the autocorrelation function $P(R)$ of the electron density. These computations are entirely analogous to the time series setting of spectral densities and autocorrelation functions.

The information in the autocorrelation function is thought to be specific for basic particle shapes [28]. Models for the particles under investigation are subsequently generated by computer, and the scattering pattern of the model is calculated with the Debye equation [6]. The experimental scattering pattern is then compared to the scattering pattern generated from the model and the model is adjusted, with the calculations repeated until the experimental data and model-derived data are similar. The model is ultimately transformed to a molecular envelope. Measurement uncertainties will likely affect the acceptance or rejection of a computer-generated model, since these uncertainties will ultimately be used in confidence intervals.

1.3 Example data for the nucleosome core particle

The nucleosome core particle consists of eight protein molecules, two copies each of the histones H2A, H2B, H3 and H4, and approximately 146 base pairs of double stranded DNA, forming $1\frac{3}{4}$ turns around the protein core. The structure of this particle has been determined by X-ray crystallography [19] and it has approximately the shape of a round disk, 110Å in diameter and 55Å thick. Small-angle X-ray scattering can help to reveal molecular structures that involve nucleosome core particles under various conditions, in presence or absence of additional protein molecules that may regulate their behavior. As an example, we show the experimental data, the Fourier inverted data and a low-resolution molecular reconstruction for the nucleosome core particle in Figure 3. The experimental data (Figure 3, left panel, gray) have been superimposed with computer-generated data (black), based on the crystal structure of the nucleosome (histone tails removed) by the program CRY SOL [27]. The superposition shows that the model adequately explains the experimental data. In the central panel, the Fourier-inverted data are shown, acquired by an automated version of GNOM [25]. Finally, the molecular envelope (for details of reconstruction, refer to Yang *et al* [35]) is shown in transparent gray, superimposed with a ribbon-model of the crystal structure (Figure 3, right). The superposition shows that the crystal structure and the molecular envelope correspond well, keeping in mind that the molecular envelope is derived from very low resolution information.

The following questions are at the basis of a statistical evaluation of SAXS data: what are the uncertainties in the recorded data, and how do they propagate into traditional (R_g) and envelope reconstruction outcome? Specifically: when can models generated by computer be accepted or rejected, based on their fit to experimental data that is presented with adequate margins of uncertainty?

1.4 Error and uncertainty analysis in SAXS

Charge-coupled devices (CCDs) were first applied to X-ray diffraction work [31] and are now in common use, including in the application of small angle X-ray scattering. One of the fundamental features of these systems is that CCDs are equipped with a phosphor or scintillation screen that captures the X-rays and converts them to visible light, which is transported to the actual CCD by a fiber optic taper. However, it has very recently been recognized in literature that the measurement uncertainties in diffraction data acquired with a CCD detector are important and not yet appropriately estimated [32]. In the case of diffraction images, localized phenomena ('spots') are recorded and compared, while in the case of X-ray scattering, these diffraction spots are not present. Instead, the image has different gray tones. The characterization of errors and uncertainties in X-ray scattering intensities is therefore more difficult than in diffraction data, yet equally important.

This paper addresses one component of the error in SAXS images, specifically the spatially autocorrelated error in the two-dimensional images. In §2.1, we propose a kernel convolution model for the random field of recorded intensity, from which we derive the pixel-to-pixel correlation structure (§2.2) in the detector plane, and annulus-to-annulus correlation structure (§2.3) for the radial averages used in subsequent analysis. We then use experimental SAXS image data from two detectors and several molecules to study the spatial correlation structure empirically (§3.1). We fit the kernel convolution model via method of moments (§3.2) and compare results from the fitted model to empirical data for one-dimensional autocorrelations (§3.3). A summary discussion follows in §4.

2 Model for correlated detector noise

2.1 Point-to-point correlation

The basic physical model of photon counting is the Poisson process, which under SAXS will have an inhomogeneous rate function that varies with the scattering angle. Photon counts over disjoint regions should behave independently, if it were not for the need to transform them into visible light and register them. This transformation and registration induces dependence, which we now describe by starting from an independence model and inducing dependence by kernel convolution, mimicking the actual physical experiment. Alternative model specifications to capture spatial correlation could include, for example, Markov random fields or multivariate autoregressive moving-average models. We do not pursue these here because they quickly depart from the basic physical model, putting correlation structure in the noise prior to kernel convolution.

Let $I_D(\mathbf{q})$ denote the *detected* intensity at each point $\mathbf{q} = (q_1, q_2)$ in the detector plane. Assume that

$$I_D(\mathbf{q}) = \mu(\mathbf{q}) + \sigma(\mathbf{q})Z(\mathbf{q}), \quad (2)$$

where $\mu(\mathbf{q})$ is the theoretical scattering intensity, $\sigma(\mathbf{q})$ is a scaling function that may depend on $\mu(\mathbf{q})$ and on detector characteristics, and $\{Z(\mathbf{q})\}$ is an independent and identically distributed random field with mean zero and variance one. In what follows, we are primarily interested in first and second-order moment properties for integrated versions of the $I_D(\cdot)$ process. On the basis of such properties, the inhomogeneous Poisson process will be largely indistinguishable from (2).

The *recorded* intensity $I_R(\mathbf{q})$ at each point $\mathbf{q}$ comes from the transfer of the detected intensity to the array of CCD pixels via fiber optic tapers, as phosphorescence occurs in the detector plane. The effect of this transfer is to spread the intensity across more than one CCD pixel,

which we model with a kernel convolution (e.g, [10, 9]). We assume that the kernel has the form

$$\frac{1}{\tau^2} K_{\xi} \left(\frac{\mathbf{s}}{\tau} \right) \quad (3)$$

where $\mathbf{s} = (s_1, s_2)$, K_{ξ} is a two-dimensional probability density function symmetric about zero, ξ is a vector of parameters, and $\tau > 0$ is a scale parameter. (The assumption of a constant scale parameter τ over space is discussed in §3.1.) We assume that τ is very small relative to the variation in $\mu(\cdot)$ and $\sigma(\cdot)$ in (2). That is, $\mu(\cdot)$ and $\sigma(\cdot)$ are approximately constant over the effective support of the kernel, which is determined by τ .

Then the recorded intensity at point $\mathbf{q}$ is given by

$$\begin{aligned} I_R(\mathbf{q}) &= \frac{1}{\tau^2} \int K_{\xi} \left(\frac{\mathbf{s}}{\tau} \right) I_D(\mathbf{q} \\ &\quad + \mathbf{s}) d\mathbf{s} \\ &= \int \frac{1}{\tau^2} K_{\xi} \left(\frac{\mathbf{s}}{\tau} \right) \mu(\mathbf{q} \\ &\quad + \mathbf{s}) d\mathbf{s} \\ &\quad + \int \frac{1}{\tau^2} K_{\xi} \left(\frac{\mathbf{s}}{\tau} \right) \sigma(\mathbf{q} \\ &\quad + \mathbf{s}) Z(\mathbf{q} \\ &\quad + \mathbf{s}) d\mathbf{s} \simeq \mu(\mathbf{q}) \\ &\quad + \sigma(\mathbf{q}) \int \frac{1}{\tau^2} K_{\xi} \left(\frac{\mathbf{s}}{\tau} \right) Z(\mathbf{q} + \mathbf{s}) d\mathbf{s} \end{aligned} \quad (4)$$

by the assumption of small τ .

The point-to-point covariance structure of recorded intensity is then given, using the small τ approximation, by

$$\text{Cov}(I_R(\mathbf{q}), I_R(\mathbf{r})) = \int \int \sigma(\mathbf{q} + \mathbf{s}) \sigma(\mathbf{r} + \mathbf{t}) \frac{1}{\tau^4} K_{\xi} \left(\frac{\mathbf{s}}{\tau} \right) K_{\xi} \left(\frac{\mathbf{t}}{\tau} \right) \mathbb{1}_{\{\mathbf{q} + \mathbf{s} = \mathbf{r} + \mathbf{t}\}} d\mathbf{s} d\mathbf{t} \simeq \sigma(\mathbf{q}) \sigma(\mathbf{r}) \int \frac{1}{\tau^4} K_{\xi} \left(\frac{\mathbf{s}}{\tau} \right) K_{\xi} \left(\frac{\mathbf{s} + \mathbf{q} - \mathbf{r}}{\tau} \right) d\mathbf{s}, \quad (5)$$

where $\mathbb{1}_{\{A\}} = 1$ if the event A is true, and 0 otherwise. From this it follows that the point-to-point correlation structure is given by

$$\text{Corr}(I_R(\mathbf{q}), I_R(\mathbf{r})) = \frac{\text{Cov}(I_R(\mathbf{q}), I_R(\mathbf{r}))}{\sqrt{\text{Var}(I_R(\mathbf{q})) \text{Var}(I_R(\mathbf{r}))}} = \frac{\int K_{\xi} \left(\frac{\mathbf{s}}{\tau} \right) K_{\xi} \left(\frac{\mathbf{s} + \mathbf{q} - \mathbf{r}}{\tau} \right) d\mathbf{s}}{\int K_{\xi}^2 \left(\frac{\mathbf{s}}{\tau} \right) d\mathbf{s}}. \quad (6)$$

These correlations are stationary in the sense of depending only on $\mathbf{q} - \mathbf{r}$, though the covariances are nonstationary due to heteroskedasticity.

The Gaussian kernel function generated from two independent $N(0, \tau^2)$ random variables,

$$\frac{1}{\tau^2} K_{\xi} \left(\frac{\mathbf{s}}{\tau} \right) = \frac{1}{2\pi\tau^2} \exp \left(\frac{-\|\mathbf{s}\|^2}{2\tau^2} \right),$$

requires no additional parameters ξ and is analytically simple and useful as we move from points to pixels to annuli. Indeed, the nice analytic properties of the Gaussian make it a

standard choice as a *point-spread function* to describe blurring in the closely-related area of microscopy (e.g., [36]). In this case, (5) becomes

$$\text{Cov}(I_R(\mathbf{q}), I_R(\mathbf{r})) \simeq \sigma(\mathbf{q})\sigma(\mathbf{r}) \frac{1}{4\pi\tau^2} \exp\left(\frac{-\|\mathbf{r} - \mathbf{q}\|^2}{4\tau^2}\right) \quad (7)$$

and (6) becomes

$$\text{Corr}(I_R(\mathbf{q}), I_R(\mathbf{r})) = \exp\left(\frac{-\|\mathbf{r} - \mathbf{q}\|^2}{4\tau^2}\right), \quad (8)$$

which is both stationary and isotropic.

2.2 Pixel-to-pixel correlation

Point-level recorded intensity is not observed. Instead, we observe the aggregate of recorded intensity over pixels. We next investigate the correlation structure of these pixel-level observations. Let $P_1 \subset \mathbf{R}^2$, $P_2 \subset \mathbf{R}^2$ denote two pixels in the detector plane. Then

$$\text{Cov}\left(\int_{P_1} I_R(\mathbf{q}) d\mathbf{q}, \int_{P_2} I_R(\mathbf{r}) d\mathbf{r}\right) \simeq \int_{P_1} \int_{P_2} \sigma(\mathbf{q})\sigma(\mathbf{r}) \int \frac{1}{\tau^4} K_\xi\left(\frac{\mathbf{s}}{\tau}\right) K_\xi\left(\frac{\mathbf{s} + \mathbf{q} - \mathbf{r}}{\tau}\right) d\mathbf{s} d\mathbf{r} d\mathbf{q} \quad (9)$$

by (5).

The ultimate detector resolution is the pixel, so it makes sense to assume that $\sigma(\mathbf{q})$ is effectively constant for $\mathbf{q} \in P_1$ and $\sigma(\mathbf{r})$ is effectively constant for $\mathbf{r} \in P_2$. Let $\sigma(\mathbf{q}^*)$, $\sigma(\mathbf{r}^*)$ denote these constants, which then cancel in the computation of the pixel-to-pixel correlation. Define

$$\Delta(\mathbf{s}; P_1, P_2) = \int_{P_1} \int_{P_2} K_\xi\left(\frac{\mathbf{s} + \mathbf{q} - \mathbf{r}}{\tau}\right) d\mathbf{r} d\mathbf{q}$$

and note that if P_2 is just a translation of the pixel P_1 , then $\Delta(\mathbf{s}; P_1, P_1) = \Delta(\mathbf{s}; P_2, P_2)$. It then follows that the pixel-to-pixel correlation is given by

$$\text{Corr}\left(\int_{P_1} I_R(\mathbf{q}) d\mathbf{q}, \int_{P_2} I_R(\mathbf{r}) d\mathbf{r}\right) \simeq \frac{\int K_\xi\left(\frac{\mathbf{s}}{\tau}\right) \Delta(\mathbf{s}; P_1, P_2) d\mathbf{s}}{\int K_\xi\left(\frac{\mathbf{s}}{\tau}\right) \Delta(\mathbf{s}; P_1, P_1) d\mathbf{s}}, \quad (10)$$

which is stationary in the sense of depending only on the pixel-to-pixel translation from P_1 to P_2 . These pixel-to-pixel correlations can be estimated from the observed data.

Consider two $w \times w$ pixels, $P_1 = (q_1^*, q_1^* + w) \times (q_2^*, q_2^* + w)$ and $P_2 = (r_1^*, r_1^* + w) \times (r_2^*, r_2^* + w)$. Let $\mathbf{q}^* = (q_1^*, q_2^*)'$ and $\mathbf{r}^* = (r_1^*, r_2^*)'$. Under the Gaussian kernel function, the pixel-to-pixel covariance can be expressed as

$$\begin{aligned} \text{Cov} \left(\int_{p_1} I_R(\mathbf{q}) d\mathbf{q}, \int_{p_2} I_R(\mathbf{r}) d\mathbf{r} \right) \\ \simeq \sigma(\mathbf{q}^*) \sigma(\mathbf{r}^*) \int_0^w \left\{ \Phi \left(\frac{r_1^* - q_1^* + w - u}{\sqrt{2\tau^2}} \right) - \Phi \left(\frac{r_1^* - q_1^* - u}{\sqrt{2\tau^2}} \right) \right\} du \\ \times \int_0^w \left\{ \Phi \left(\frac{r_2^* - q_2^* + w - u}{\sqrt{2\tau^2}} \right) - \Phi \left(\frac{r_2^* - q_2^* - u}{\sqrt{2\tau^2}} \right) \right\} du \end{aligned} \quad (11)$$

where $\Phi(\cdot)$ is the cumulative distribution function of a standard normal random variable. Let $(h_1, h_2) = (r_1^* - q_1^*, r_2^* - q_2^*)$. The pixel-to-pixel correlation is then

$$\rho(h_1, h_2) = \frac{\int_0^w \left\{ \Phi \left(\frac{h_1 + w - u}{\sqrt{2\tau^2}} \right) - \Phi \left(\frac{h_1 - u}{\sqrt{2\tau^2}} \right) \right\} du \int_0^w \left\{ \Phi \left(\frac{h_2 + w - u}{\sqrt{2\tau^2}} \right) - \Phi \left(\frac{h_2 - u}{\sqrt{2\tau^2}} \right) \right\} du}{\int_0^w \left\{ \Phi \left(\frac{w - u}{\sqrt{2\tau^2}} \right) - \Phi \left(\frac{-u}{\sqrt{2\tau^2}} \right) \right\} du \int_0^w \left\{ \Phi \left(\frac{w - u}{\sqrt{2\tau^2}} \right) - \Phi \left(\frac{-u}{\sqrt{2\tau^2}} \right) \right\} du}, \quad (12)$$

which we evaluate via numerical integration. Similar expressions can be derived for other kernel functions formed as the product of identical probability densities in each dimension.

For any such product kernel, like the Gaussian kernel considered here, $\rho(h, 0) = \rho(-h, 0)$, $\rho(0, h) = \rho(0, -h)$ and horizontal autocorrelations equal vertical autocorrelations:

$$\rho(h, 0) = \rho(0, h).$$

From this it follows that

$$\rho(h_1, h_2) = \rho(h_1, 0) \rho(0, h_2).$$

In particular, for $t = 0, \pm 1, \pm 2, \dots$, we have “main diagonal” pixels with lower-left corners $(q_1^* + tw, q_2^* - tw)$ and autocorrelations

$$\rho(tw, -tw) = \rho(tw, 0) \rho(0, -tw) = \rho^2(tw, 0),$$

and “minor diagonal” pixels with lower-left corners $(q_1^* + tw, q_2^* + tw)$ and autocorrelations

$$\rho(tw, tw) = \rho(tw, 0) \rho(0, tw) = \rho^2(tw, 0).$$

2.3 Annulus-to-annulus correlation

SAXS images are typically reduced to one-dimensional curves via radial averaging, as depicted in Figure 2. In practice, this means assigning pixels to a particular angle by some rule. The rule effectively approximates an annulus with circular borders by a set of pixels with piecewise linear borders. The radial average is computed as the mean intensity for the assigned pixels. Let α_1 denote the set of pixels assigned to angle 1, and α_2 the set of pixels assigned to angle 2. Then the covariance between the radial averages would be

$$\frac{1}{|\alpha_1| |\alpha_2|} \sum_{k \in \alpha_1} \sum_{\ell \in \alpha_2} \int_{p_k} \int_{p_\ell} \sigma(\mathbf{q}) \sigma(\mathbf{r}) \int \frac{1}{\tau^4} K_\xi \left(\frac{\mathbf{s}}{\tau} \right) K_\xi \left(\frac{\mathbf{s} + \mathbf{q} - \mathbf{r}}{\tau} \right) d\mathbf{s} d\mathbf{r} d\mathbf{q}$$

by (9).

For a given assignment rule and kernel function, this covariance and the corresponding correlation are readily computed. But the finite sums make these expressions not particularly amenable to further analytic study, so we now consider the correlation structure of radial averages over annuli (with circular borders) in the detector plane. Let each annulus be of width w , and convert to polar coordinates:

$$A_1 = \{(q \cos(\phi), q \sin(\phi)) : q^* < |q| \leq q^* + w, 0 \leq \phi < 2\pi\}$$

and

$$A_2 = \{(r \cos(\psi), r \sin(\psi)) : r^* < |r| \leq r^* + w, 0 \leq \psi < 2\pi\}.$$

Then the covariance of the radial averages is given by

$$\begin{aligned} \text{Cov} \left(\frac{1}{|A_1|} \int_{A_1} I_R(q \cos(\phi), q \sin(\phi)) dq d\phi, \frac{1}{|A_2|} \int_{A_2} I_R(r \cos(\psi), r \sin(\psi)) dr d\psi \right) \\ \simeq \frac{1}{4\pi^2 w^2} \int_{A_1} \int_{A_2} \sigma(\mathbf{q}) \sigma(\mathbf{r}) \int \frac{1}{\tau^4} K_\xi \left(\frac{\mathbf{s}}{\tau} \right) K_\xi \left(\frac{\mathbf{s} + \mathbf{q} - \mathbf{r}}{\tau} \right) ds dr d\psi dq d\phi \end{aligned}$$

by (5).

The theoretical intensity is assumed radially symmetric, so for narrow annuli we assume that $\sigma(\mathbf{q})$ is constant for $\mathbf{q} \in A_1$ and $\sigma(\mathbf{r})$ is constant for $\mathbf{r} \in A_2$. Let $\sigma(\mathbf{q}^*)$, $\sigma(\mathbf{r}^*)$ denote these constants, which then cancel in the computation of the annulus-to-annulus correlation. Define

$$\Lambda(\mathbf{s}; A_1, A_2) = \int_{A_1} \int_{A_2} K_\xi \left(\frac{\mathbf{s} + (q \cos(\phi) - r \cos(\psi), q \sin(\phi) - r \sin(\psi))}{\tau} \right) dr d\psi dq d\phi.$$

It then follows that the annulus-to-annulus correlation is given by

$$\text{Corr} \left(\frac{1}{|A_1|} \int_{A_1} I_R(\mathbf{q}) d\mathbf{q}, \frac{1}{|A_2|} \int_{A_2} I_R(\mathbf{r}) d\mathbf{r} \right) \simeq \frac{\int K_\xi \left(\frac{\mathbf{s}}{\tau} \right) \Lambda(\mathbf{s}; A_1, A_2) d\mathbf{s}}{\left\{ \int K_\xi \left(\frac{\mathbf{s}}{\tau} \right) \Lambda(\mathbf{s}; A_1, A_1) d\mathbf{s} \int K_\xi \left(\frac{\mathbf{s}}{\tau} \right) \Lambda(\mathbf{s}; A_2, A_2) d\mathbf{s} \right\}^{1/2}}. \quad (13)$$

Under the Gaussian kernel function, the annulus-to-annulus covariance can be expressed as

$$\begin{aligned} \text{Cov} \left(\frac{1}{|A_1|} \int_{A_1} I_R(q \cos(\phi), q \sin(\phi)) dq d\phi, \frac{1}{|A_2|} \int_{A_2} I_R(r \cos(\psi), r \sin(\psi)) dr d\psi \right) \\ \simeq \frac{\sigma(\mathbf{q}^*) \sigma(\mathbf{r}^*)}{4\pi^2 w^2 \sqrt{4\pi\tau^2}} \int_{q^*}^{q^*+w} \int_0^{2\pi} \int_0^{2\pi} \exp \left(\frac{-q^2 \sin^2(\phi - \psi)}{4\tau^2} \right) \\ \times \left\{ \Phi \left(\frac{r^* + w - q \cos(\phi - \psi)}{\sqrt{2}\tau^2} \right) - \Phi \left(\frac{r^* - q \cos(\phi - \psi)}{\sqrt{2}\tau^2} \right) \right\} d\phi d\psi dq \end{aligned} \quad (14)$$

$$= \frac{\sigma(q^*)\sigma(r^*)}{4\pi^2 w^2 \sqrt{4\pi\tau^2}} G(q^*, r^*, w). \quad (15)$$

The annulus-to-annulus correlation is then

$$\lambda(q^*, r^*, w) = \frac{G(q^*, r^*, w)}{\sqrt{G(q^*, q^*, w)G(r^*, r^*, w)}}, \quad (16)$$

which is not stationary.

3 Empirical results

3.1 Spatial autocorrelations from SAXS experiments

We investigated spatial autocorrelation across a range of experimental conditions, including data from two different beam lines (ALS 12.3.1 and SSRL 4-2) and several sample types: no macromolecule (buffer only), protein (glucose isomerase), DNA, and a protein-DNA complex. We evaluated these fundamentally different samples with the question in mind whether the physical sample characteristics might affect the scattering behavior and spatial autocorrelation results. Proteins are different from nucleic acids (DNA, RNA) mostly by virtue of the presence of phosphates in nucleic acids. Compounds containing phosphorus atoms, due to their increased atomic scattering power, give a relatively stronger scattering pattern, compared to compounds that contain only C, O, N and H atoms (contributing to protein scattering). Finally, for each macromolecular SAXS experiment, an experiment with the identical buffer, without macromolecule, is performed for subsequent data subtraction and we therefore considered this class of samples separately. The exact samples used, their origin, preparation method and experimental treatment are given in the appendix.

To estimate spatial autocorrelation, we first removed mean structure by subtracting two replicate images. Because there are many sources of variation from image to image, including variation in buffer, macromolecule, dilution of macromolecule in buffer, X-ray source intensity and exposure time, we matched images that were collected on identical samples at the same dilution, and as close in time as possible. We compared subsequent images from the same sample (previously exposed) with images from fresh samples. For the samples chosen for this work the best image-to-image match was found for images derived from a repeated exposure of the same sample (i.e. radiation damage did not occur).

Inspection of the resulting mean-corrected images shows that most of the mean structure ($\mu(\cdot)$ in (4)) is in fact gone, though there may be local sources of variation around the shadow of the beam stop and the edges of the image. The heteroskedasticity described by $\sigma(\cdot)$ is also present. To correct for local mean or variance structure, we divide the detector plane into short “time series” of adjacent pixels, horizontally or vertically or diagonally, and look at the empirical autocorrelations of these local series. We chose $n = 32$ for the length of our series, because it is small relative to the size of the detector plane, yet sufficiently large for fairly stable estimation of low-lag autocorrelations. It is also a common factor of the horizontal and vertical dimension of the two detectors we considered, so that we get an integer number of disjoint series.

For each differenced image, we then computed all the empirical autocorrelations at several lags. For each differenced image, each direction (horizontal, vertical, main diagonal, minor diagonal), and each lag, we produced a spatial image (not shown) of the set of empirical autocorrelations. These images show no discernible patterns, indicating that a locally stationary and isotropic model for the errors is appropriate, with nonstationarity due only to

local mean effects (like the shadow of the beam stop and the edges) and heteroskedasticity. In particular, this empirical evidence suggests that a single kernel with constant scale parameter τ suffices to model the spatial autocorrelation structure.

In Figures 4–5, we show summary results for all of the estimated autocorrelations at lags one through three, across the entire detector. Subsequent lags look like lag three and are not shown. The asymptotic Bartlett bounds at $\pm 1.96n^{-1/2}$ (e.g., [1], Theorem 7.2.1) are shown along with a reference line at zero. There is remarkable consistency across all of the experimental conditions. Within each experiment, the horizontal and vertical autocorrelations are nearly identical, as are the main and minor diagonal autocorrelations. With the exception of lag one, the boxplots are centered slightly below the zero reference line. (This is consistent with the negative bias of the empirical autocorrelations in small samples, as indicated by the “Simulation” boxplots.) Most of the empirical autocorrelations are within the Bartlett bounds. Only at lag one in each direction are the exceedances of the Bartlett bounds greater than the nominal 5% expected under an uncorrelated model. At lags two and three, the exceedances are close to those predicted from the asymptotic theory.

3.2 Method of moments estimation of kernel convolution model

We fitted the pixel-to-pixel correlation structure using the Gaussian kernel function, estimating the single parameter, τ , for each molecule type via method of moments. First, we evaluated $E_{\tau} [\hat{\rho}(1,0)]$ via simulation by generating 10,000 independent Gaussian time series of length 32, each with the appropriate covariance structure given τ , and computing the average empirical autocorrelation across all 10,000 replications. This expectation is not equal to $\rho(1,0)$, due to the small-sample bias of the empirical autocorrelations.

We repeated this simulation on a grid of τ values from 0.37 to 0.42, establishing that $E_{\tau} [\hat{\rho}(1,0)]$ is linear to an excellent approximation over this range. Specifically,

$$\tau = 0.1264 + 0.9074 E_{\tau} [\hat{\rho}(1,0)] \quad (17)$$

(0.00010) (0.00035)

where the numbers in parentheses indicate standard errors from ordinary least squares regression, and $R^2 > 0.9999$ for the fit. The same equation applies to vertical autocorrelation, $E_{\tau} [\hat{\rho}(0,1)]$.

Similarly, we established that $E_{\tau} [\hat{\rho}(1,1)]$ is linear to an excellent approximation over the same range of τ values,

$$\tau = 0.2848 + 1.3237 E_{\tau} [\hat{\rho}(1,1)], \quad (18)$$

(0.00039) (0.00459)

with $R^2 = 0.9997$ for the fit.

One method of moments estimator (MOME) was then computed for each direction, with “Horizontal” and “Vertical” MOME’s obtained from (17) via

$$\hat{\tau}_{\text{Horiz}} = 0.1264 + 0.9074 \overline{\hat{\rho}(1,0)}, \hat{\tau}_{\text{Vert}} = 0.1264 + 0.9074 \overline{\hat{\rho}(0,1)},$$

where $\overline{\hat{\rho}(1,0)}$ and $\overline{\hat{\rho}(0,1)}$ are the estimated lag-one horizontal and vertical autocorrelations, averaged across all relevant time series from the differenced images. Similarly, “Main” diagonal and “Minor” diagonal MOME’s were obtained from (18) via

$$\widehat{\tau}_{\text{Main}} = 0.2848 + 1.3237 \overline{\widehat{\rho}(1, -1)}, \widehat{\tau}_{\text{Minor}} = 0.2848 + 1.3237 \overline{\widehat{\rho}(1, 1)}.$$

These four MOME's are tabled for each molecule in Table 1.

For the horizontal direction, the standard error of $\hat{\tau}$ is

$$\text{SE}(\widehat{\tau}) = 0.9074 \times \text{SE}(\overline{\widehat{\rho}(1, 0)}),$$

where $\text{SE}(\overline{\widehat{\rho}(1, 0)})$ was evaluated via a stratified block bootstrap method (e.g., [7, 3, 15, 21]), to account for spatial variation across the detector and spatial dependence among the estimated autocorrelations. For the 512×512 detector, we formed 16×16 strata of 32 horizontal time series, each of length 32. Within each stratum, we formed 8 disjoint horizontal blocks, each consisting of 4 adjacent horizontal time series. We then drew 8 blocks within each stratum via simple random sampling with replacement. The average of the empirical lag-one horizontal autocorrelations across the entire stratified sample (32 autocorrelations in each of 16×16 strata) yielded the b th bootstrap replicate estimate,

$$\overline{\widehat{\rho}(1, 0)}^{(b)},$$

for $b = 1, 2, \dots, B$. We chose $B = 100$. The bootstrap standard error is then

$$\text{SE}(\overline{\widehat{\rho}(1, 0)}) = \left\{ \frac{\sum_{b=1}^B (\overline{\widehat{\rho}(1, 0)}^{(b)} - B^{-1} \sum_{b=1}^B \overline{\widehat{\rho}(1, 0)}^{(b)})^2}{B-1} \right\}^{1/2}.$$

We used the same approach for the MOME based on the vertical autocorrelations, and for the other detector (using 24×24 strata). These bootstrap standard errors are given in Table 1.

We also computed approximate standard errors for each MOME via an asymptotic approximation. We used $\hat{\tau}$ to compute the autocorrelation function of a stationary time series in the given direction (horizontal/vertical or diagonal), then used Bartlett's formula ([1], Theorem 7.2.1) to compute the asymptotic variance of the lag-one autocorrelation for such a series of length 32, $w_{\hat{\tau}}/32$. We then treated the estimated lag-one autocorrelations for disjoint series of length 32 as uncorrelated, so that the average lag-one autocorrelation has approximate standard error

$$n^{-1/2} \sqrt{w_{\hat{\tau}}/32},$$

where n is the number of empirical autocorrelations in the average. This "Bartlett standard error" is given in Table 1, along with the sample sizes, n . There is very good agreement between the bootstrap standard errors and the Bartlett standard errors in the vertical/horizontal directions.

There is excellent agreement between $\hat{\tau}_{\text{vert}}$ and $\hat{\tau}_{\text{horiz}}$, even given the very small standard errors. The estimates based on the main and minor diagonal autocorrelations also appear quite close to one another. There appears to be a statistically significant discrepancy in each case between the estimates based on the horizontal/vertical autocorrelations and those based on the diagonal autocorrelations. While this indicates lack of fit of the simple kernel model, the lack of fit is remarkably small for a one-parameter model, given the amount of data available.

As a simple diagnostic, we simulated 8,192 independent realizations from a Gaussian model with autocorrelations given by (12) with $\tau = 0.39$. For these realizations, we computed empirical autocorrelations at lags 1–3 for horizontal, vertical, and diagonal sequences of $n = 32$ pixels each, and plotted the corresponding boxplots in Figures 4–5. There is quite good, but not perfect, agreement between the empirical boxplots and the simulation boxplots under the kernel convolution model. In particular, the lag-one simulation boxplot in the horizontal or vertical dimension has less variation than the corresponding empirical boxplots. Possible explanations for this include 1) the simulated series were all independent, but the empirical series are not independent; 2) the simulated series used Gaussian innovations, while the real data may have heavier tails; 3) the simulation is entirely stationary, while the empirical autocorrelations include effects of edges and the beam stop shadow; and 4) the simulation is entirely stationary, while the real data may have varying degrees of autocorrelation on different parts of the detector. Once again, evidence of lack of fit is not surprising given the large amount of data at hand; what is remarkable is that a simple, one-parameter model fits quite well and consistently across a very broad range of experimental conditions.

3.3 Correlations for one-dimensional curves

We computed the annulus-to-annulus correlations for $w = 0.5$ pixels and $\tau = 0.39$, which are consistent with data from the empirical study. In Figure 6, we plot these theoretical autocorrelations for $r^* = 10, 200$ or 500 and for q^* in a neighborhood of r^* . The autocorrelations are non-stationary, since the curves vary with r^* , and anisotropic, since the curves are asymmetric. The correlations are quite different from zero at lags one and two and are nearly zero thereafter. While these correlations are for the radial averages over circular annuli, not for radial averages over the actual pixelated approximations to annuli (see §2.3), the theoretical correlations give some indication of what range of dependence might be expected in real data.

We next considered a set of 14 one-dimensional curves for glucose isomerase, $\{X_{it}\}$ for $i = 1, \dots, 14$ and $t = 1, \dots, 500$. These were all computed as replicates from a common source, with the same exposure time and with a common buffer image. We removed the mean structure by subtracting curve i from curve $i+1$, $\{Y_{it}\} = \{X_{i,t+1} - X_{it}\}$ for $i = 1, \dots, 13$. We then looked at disjoint sequences of length 100 from the resulting series, $\{Y_{it}\}_{t=1}^{100}, \dots, \{Y_{it}\}_{t=401}^{500}$. Four of these 65 sequences with corresponding ACF's are shown in Figure 7. In each case, the difference looks stationary and standard identification techniques would lead to a second-order moving average, MA(2). This is true for 37 of the 65 sequences. Of those remaining, 8 look like other low-order MA processes and 20 appear non-stationary. The 20 all correspond to indices 1–100 (13 cases) or 101–200 (7 cases). The explanation for the non-stationary behavior is not clear, but indicates replicate-to-replicate differences in mean structure at low angles, possibly corresponding to sampling variation from replicate to replicate. Such potential errors are a source of ongoing investigation. Computing the correct correlation structure for radial averages would require knowing the exact algorithm for matching pixels to annuli, which is information we do not have. In spite of these limitations, it is interesting to note that at least the range of correlation dependence in the empirical one-dimensional curves is consistent with the fitted kernel convolution model.

4 Discussion

There are a number of potential implications of autocorrelation in the image and in the radial averages. Estimates of physical characteristics like radius of gyration, maximum linear dimension of a particle in solution and particle volume rely on parametric and semiparametric regression methods. Efficient estimation in the parametric case, effective tuning parameter selection in the semiparametric case, and appropriate standard errors and confidence intervals in either case rely on correct accounting of the dependence structure in the errors. To the best of our knowledge, this has not been done for SAXS data. Further, any attempt to quantify the quality of a shape reconstruction for a biological macromolecule, or to compare the quality of different shape reconstructions, will need to account for sources of error and uncertainty in the original detector data, including the autocorrelated errors identified here.

Acknowledgments

The work of Erciulescu was conducted while she was a student in the Departments of Mathematics and Statistics at Colorado State University. This research was supported by the Joint NSF/NIGMS Initiative to Support Research in the Area of Mathematical Biology (R01GM096192).

Dr. Joseph Ferrara (Rigaku Americas) and Dr. Henry Bellamy (Louisiana State University) are acknowledged for useful discussions.

X-ray scattering and diffraction technologies and their applications to the determination of macromolecular shapes and conformations at the SIBYLS beamline at the Advanced Light Source, Lawrence Berkeley National Laboratory are supported in part by the DOE program Integrated Diffraction Analysis Technologies (IDAT) and the DOE program Molecular Assemblies Genes and Genomics Integrated Efficiently (MAGGIE) under Contract Number DE-AC02-05CH11231 with the U.S. Department of Energy; Efforts to apply SAXS and crystallography to characterize eukaryotic pathways relevant to human cancers are supported in part by National Cancer Institute grant CA92584.

SAXS data were collected in part at the Stanford Synchrotron Radiation Laboratory, a national user facility operated by Stanford University on behalf of the U.S. Department of Energy, Office of Basic Energy Sciences. The SSRL Structural Molecular Biology Program is supported by the Department of Energy, Office of Biological and Environmental Research, and by the National Institutes of Health, National Center for Research Resources, Biomedical Technology Program.

Appendix: experimental methods and initial data processing

Glucose isomerase was acquired from Hampton Research (Aliso Viejo, CA) and dialyzed extensively against a buffer of 10mM Tris-HCl, pH 8.0 and 100mM NaCl. The preparation and data collection on the DNA used in this experiment (blunt 30-mer) was previously described [16]. Nucleosomes were prepared as previously described [4], using 146 base pair DNA fragments derived from the strong positioning '601' DNA sequence [17]. The experimental nucleosome treatment, data collection and processing are identical to those described by Yang *et al* [35].

SAXS data for most samples were collected at the SIBYLS beam line (12.3.1) at the ALS (Berkeley) with an X-ray energy of 10 keV ($\eta = 1.2398\text{\AA}$). A Mar CCD detector was used to record the scattering data. A 15 μl sample was placed in a 1mm thick chamber with two windows of 25 μm mica. The distance between the sample and the detector was 1.5m. Stock solutions of glucose isomerase, DNA or nucleosomes were each used at full, $\frac{2}{3}$ and $\frac{1}{3}$ concentrations. The raw images (corrected with dark images, but reference images not subtracted) were used for statistical analysis.

In short, data derived from scattering images for biophysical interpretation were processed as follows. After image subtraction and radial integration, initial data processing was

performed with the program PRIMUS [13]. The radius of gyration, R_g , was estimated by Guinier analysis [6] in PRIMUS using low-angle data ($sR_g < 1.3$); the maximum particle dimension was estimated by regularized indirect Fourier transform with the program GNOM [24], which provides the $P(r)$ function. The $P(r)$ function is a histogram of inter-atomic distances and goes to zero at the maximum intramolecular distance d_{max} .

The molecular envelope was made by taking ten low-resolution molecular envelopes calculated from the scattering data for each model, using the program DAMMIN [26] and superimposed using SUPCOMB [14]. Finally, the pre-aligned models were averaged with DAMAVER [30], giving an effective occupancy of each voxel. A convex shell of all models can be generated by keeping all occupied voxels, while filtering at half-maximal occupancy provides “filtered” models. To visualize the results, the reconstructed models were converted to volumetric maps using real space convolution with a Gaussian kernel with the program Situs [33, 34]. A kernel width of 6 Å and voxel spacing of 1 Å were used. Molecular representations were made with VMD [11].

References

1. Brockwell, PJ.; Davis, RA. Time Series: Theory and Methods. 2nd edition. New York: Springer-Verlag; 1991.
2. Cantor, CR.; Schimmel, PR. Biophysical Chemistry, Part II: Techniques for the study of biological structure and function. Vol. volume II. W.H. Freeman and Company; 1980.
3. Carlstein E. The use of subseries values for estimating the variance of a general statistic from a stationary sequence. The Annals of Statistics. 1986; 14(3):1171–1179.
4. Dyer, PN.; Edayathumangalam, RS.; White, CL.; Bao, Y.; Chakravarthy, S.; Muthurajan, UM.; Luger, K. Reconstitution of nucleosome core particles from recombinant histones and DNA. In: Allis, C.; Wu, CD., editors. Methods in Enzymology. Vol. volume 375. Elsevier Academic Press; 2004. p. 23-43.
5. Feigin, L.; Svergun, D. Structure Analysis by Small-Angle X-ray and Neutron Scattering. Plenum Press; 1987.
6. Guinier, A.; Fournet, G. Small Angle Scattering of X-rays. New York: Wiley Interscience; 1955.
7. Hall P. Resampling a coverage pattern. Stochastic Processes and Their Applications. 1985; 20(2): 231–246.
8. Hansen JC. Conformational dynamics of the chromatin fiber in solution: Determinants, mechanism, and functions. Annual Review in Biophysical and Biomolecular Structure. 2002; 31:361–392.
9. Higdon, D. Quantitative Methods for Current Environmental Issues. London: Springer-Verlag; 2002. Space and space-time modeling using process convolutions; p. 37-56.
10. Higdon, D.; Swall, J.; Kern, J. Non-stationary spatial modeling. In: Bernardo, JM.; Berger, JO.; Dawid, AP.; Smith, AFM., editors. Bayesian Statistics 6. Oxford: Oxford University Press; 1998. p. 761-768.
11. Humphrey W, Dalke A, Schulten K. VMD: Visual molecular dynamics. Journal of Molecular Graphics. 1996; 14(1):33–38. [PubMed: 8744570]
12. H. G. S. C. International. Finishing the euchromatic sequence of the human genome. Nature. 2004 Oct; 431(7011):931–945. [PubMed: 15496913]
13. Konarev P, Volkov V, Sokolova A, Koch M, Svergun D. PRIMUS: a windows pc-based system for small-angle scattering data analysis. Journal of Applied Crystallography. 2003; 36:1277–1282.
14. Kozin M, Svergun D. Automated matching of high- and low-resolution structural models. Journal of Applied Crystallography. 2001; 34:33–41.
15. Künsch H. The jackknife and the bootstrap for general stationary observations. The Annals of Statistics. 1989; 17(3):1217–1241.
16. Lilyestrom W, van der Woerd MJ, Clark N, Luger K. Structural and biophysical studies of human PARP-1 in complex with damaged DNA. Journal of Molecular Biology. 2010; 395(5):983–994. [PubMed: 19962992]

17. Lowary PT, Widom J. New DNA sequence rules for high affinity binding to histone octamer and sequence-directed nucleosome positioning. *Journal of Molecular Biology*. 1998; 276(1):19–42. [PubMed: 9514715]
18. Luger K, Hansen JC. Nucleosome and chromatin fiber dynamics. *Current Opinion in Structural Biology*. 2005; 15:188–196. [PubMed: 15837178]
19. Luger K, Mäder AW, Richmond RK, Sargent DF, Richmond TJ. Crystal structure of the nucleosome core particle at 2.8Å resolution. *Nature*. 1997 Sep; 389(6648):251–260. [PubMed: 9305837]
20. Nikitina T, Ghosh RP, Horowitz-Scherer RA, Hansen JC, Grigoryev SA, Woodcock CL. MeCP2-chromatin interactions include the formation of chromatosome-like structures and are altered in mutations causing Rett syndrome. *Journal of Biological Chemistry*. 2007; 282:28237–28245. [PubMed: 17660293]
21. Politis D. The impact of bootstrap methods on time series analysis. *Statistical Science*. 2003; 18(2): 219–230.
22. Porod, G. General theory. In: Galtter, O.; Kratky, O., editors. *Small Angle X-ray Scattering*. Academic Press; 1982.
23. Putnam CD, Hammel M, Hura GL, Tainer JA. X-ray solution scattering (SAXS) combined with crystallography and computation: defining accurate macromolecular structures, conformations and assemblies in solution. *Quarterly Reviews of Biophysics*. 2007; 40(3):191–285. [PubMed: 18078545]
24. Semenyuk A, Svergun D. GNOM - a program package for small-angle scattering data-processing. *Journal of Applied Crystallography*. 1991; 24:537–540.
25. Svergun DI. Determination of the regularization parameter in indirect-transform methods using perceptual criteria. *Journal of Applied Crystallography*. 1992; 25:495–503.
26. Svergun DI. Restoring low resolution structure of biological macromolecules from solution scattering using simulated annealing. *Biophysical Journal*. 1999; 76(6):2879–2886. [PubMed: 10354416]
27. Svergun DI, Barberato CC, Koch M. CRY SOL: a program to evaluate X-ray solution scattering of biological macromolecules from atomic coordinates. *Journal of Applied Crystallography*. 1995; 28:768–773.
28. Svergun DI, Koch M. Small-angle scattering studies of biological macromolecules in solution. *Reports on Progress in Physics*. 2003; 66(10):1735–1782.
29. Thomas JO. Histone H1: location and role. *Current Opinion in Cell Biology*. 1999; 11(3):312–317. [PubMed: 10395563]
30. Volkov VV, Svergun DI. Uniqueness of *ab initio* shape determination in small-angle scattering. *Journal of Applied Crystallography*. 2003; 36:860–864.
31. Walter RL, Thiel DJ, Barna SL, Tate MW, Wall ME, Eikenberry EF, Gruner SM, Ealick SE. High-resolution macromolecular structure determination using CCD detectors and synchrotron radiation. *Structure*. 1993; 3(8):835–844. [PubMed: 7582900]
32. Waterman D, Evans G. Estimation of errors in diffraction data measured by CCD area detectors. *Journal of Applied Crystallography*. 2010 Dec; 43(6):1356–1371.
33. Wriggers W, Chacón P. Using Situs for the registration of protein structures with low-resolution bead models from x-ray solution scattering. *Journal of Applied Crystallography*. 2001; 34:773–776.
34. Wriggers W, Milligan RA, McCammon JA. Situs: A package for docking crystal structures into low-resolution maps from electron microscopy. *Journal of Structural Biology*. 1999; 125:185–195. [PubMed: 10222274]
35. Yang C, van der Woerd MJ, Muthurajan UM, Hansen JC, Luger K. Biophysical analysis and small-angle x-ray scattering-derived structures of MeCP2-nucleosome complexes. *Nucleic Acids Research*. 2011; 39:4122–4135. [PubMed: 21278419]
36. Zhang B, Zerubia J, Olivo-Marin J. Gaussian approximations of fluorescence microscope point-spread function models. *Applied Optics*. 2007; 46(10):1819–1829. [PubMed: 17356626]

Figure 1. Schematic illustration of a SAXS experiment. The sample (center) is exposed to an incident X-ray beam and the resulting scattering pattern is recorded on a two-dimensional detector. Not shown is the beamstop (which prevents the X-ray beam from hitting the center of the detector directly) and its holder. A typical scattering pattern is shown in Figure 2.

Figure 2.

(Left) Two-dimensional X-ray scattering image acquired from a 3mg/ml nucleosome core particle solution. The white feature is the shadow of the beamstop (which prevents the X-ray beam from hitting the detector directly) and its holder. Averaging of the pixels in a radial pattern as indicated by the superimposed arrow gives a one-dimensional pattern. The conversion from detector position (the center is the reference point) to $S = 4\pi \sin(\theta)/\eta$ with θ the angle between incident and scattered beams is established by a calibration sample, where η is the wavelength of the incident radiation. (Right) One-dimensional curve of intensity (logarithmic scale) versus S . The intensity (I) values are the average over pixels in concentric annuli and the error bars the standard deviations calculated using these points. The arrow indicates the data point that is the result from the averaging process in the left-hand panel.

Figure 3.

SAXS scattering data, derived physical parameters, and shape reconstructions. (Left) SAXS scattering data for the nucleosome core particle (gray) superimposed with data generated from the crystal structure (Protein Data Bank entry 1AOI; scattering data calculated by program CRY SOL, in black). (Middle) Indirect inverse Fourier transformation of the experimental data gives a curve returning to zero at 110 Å. This value is the maximum linear dimension of the particle, in good agreement with the crystal structure. (Right) Two perpendicular views of the reconstructed envelope (transparent gray) superimposed with a ribbon-model made from the crystal structure (black). The envelope and the model superimpose well.

Autocorrelation Boxplots Horizontal and Vertical Neighbors

Figure 4.

Boxplots summarizing lag 1–3 empirical autocorrelations for all (across the entire detector) horizontal and vertical sequences of $n = 32$ pixels. Light gray represents the horizontal direction and dark gray represents vertical. Dashed vertical reference lines are the Bartlett bounds, $\pm 1.96n^{-1/2}$; approximately 95% of the empirical autocorrelations should be inside the Bartlett bounds for uncorrelated sequences. Data for Buffer, Protein1, DNA, and Protein-DNA1 all come from 512×512 pixel images acquired at the SIBYLS beam line (12.3.1) at the ALS (Berkeley). The remaining boxplots come from 768×768 pixel images acquired at the SSRL 4–2 beam line (Stanford). See Table 1 for further details. Simulation boxplots, shown with cross-hatching, are based on 8,192 independent realizations from a Gaussian

model with autocorrelations given by (12) with $\tau = 0.39$. The boxplots demonstrate that (i) significant positive spatial autocorrelation is present at lag 1, with weak negative autocorrelation at lags 2 and 3; (ii) these autocorrelations are consistent across experimental samples and detectors; and (iii) these correlations are described well by the kernel convolution model. In particular, the weak negative autocorrelation at lags 2 and 3 is consistent with the simulation, showing that it results from small sample bias of the empirical autocorrelations.

\$watermark-text

\$watermark-text

\$watermark-text

Autocorrelation Boxplots Main and Minor Diagonal Neighbors

Figure 5.

Boxplots summarizing lag 1–3 empirical autocorrelations for all (across the entire detector) main and minor diagonal sequences of $n = 32$ pixels. Light gray represents the main diagonal direction and dark gray represents minor diagonal. Simulation boxplots, shown with cross-hatching, are based on 8,192 independent realizations from a Gaussian model with autocorrelations given by (12) with $\tau = 0.39$. The boxplots demonstrate that (i) significant positive spatial autocorrelation is present at lag 1, with weak negative autocorrelation at lags 2 and 3; (ii) these autocorrelations are consistent across experimental samples and detectors; and (iii) these correlations are described well by the kernel convolution model. In particular, the weak negative autocorrelation at lags 2 and 3 is

consistent with the simulation, showing that it results from small sample bias of the empirical autocorrelations.

\$watermark-text

\$watermark-text

\$watermark-text

Figure 6.

Autocorrelations computed from (16) for annuli under the kernel convolution model with $\tau = 0.39$, $w = 0.5$ and $r^* = 10$ (black 1's), $r^* = 200$ (medium gray 2's) or $r^* = 500$ (light gray 5's). Non-stationarity is indicated by the different curves for different values of r^* .

Anisotropy is indicated by the asymmetry of the curves. Note also the drop-off to near zero correlation beyond lag two.

Figure 7.

Four examples of the difference between replicate radial average curves for glucose isomerase, with 100 angles each, and corresponding sample autocorrelation functions with dashed reference lines for statistical significance at the Bartlett bounds ± 0.196 . Differences appear stationary with autocorrelation structure that is non-zero for two lags and near zero thereafter.

Table 1

Average lag-one autocorrelations from horizontal, vertical, main diagonal, or minor diagonal directions and corresponding method-of-moments estimates for τ . Sample size is number of estimated lag-one autocorrelations in the average. Bootstrap standard errors for $\hat{\tau}$ in horizontal and vertical directions are obtained via stratified block bootstrap. Bartlett standard errors are asymptotic approximations.

Molecule	Direction	Lag-one Avg. ACF	Sample Size	$\hat{\tau}$	Bootstrap SE	Bartlett SE
Buffer	Vert.	0.2913	6788	0.3908	0.0025	0.0023
	Horiz.	0.2903	6788	0.3898	0.0020	0.0023
	Main	0.0976	6681	0.4141		0.0030
	Minor	0.0976	6681	0.4140		0.0030
Protein1	Vert.	0.2898	6788	0.3894	0.0017	0.0023
	Horiz.	0.2881	6788	0.3878	0.0018	0.0023
	Main	0.0974	6681	0.4138		0.0030
	Minor	0.1000	6681	0.4172		0.0030
Protein2	Vert.	0.2966	18426	0.3955	0.0013	0.0014
	Horiz.	0.2943	18424	0.3934	0.0013	0.0014
	Main	0.0911	17664	0.4055		0.0018
	Minor	0.0923	17664	0.4070		0.0018
DNA	Vert.	0.2781	6788	0.3787	0.0018	0.0023
	Horiz.	0.2851	6788	0.3851	0.0025	0.0023
	Main	0.0876	6681	0.4007		0.0029
	Minor	0.0890	6680	0.4026		0.0029
Protein-DNA1	Vert.	0.2713	6788	0.3725	0.0023	0.0023
	Horiz.	0.2739	6788	0.3749	0.0023	0.0023
	Main	0.0799	6681	0.3905		0.0029
	Minor	0.0813	6680	0.3924		0.0029
Protein-DNA2	Vert.	0.2954	18425	0.3944	0.0013	0.0014
	Horiz.	0.2934	18425	0.3926	0.0011	0.0014
	Main	0.0908	17664	0.4050		0.0018
	Minor	0.0940	17664	0.4092		0.0018