Wilson Statistics: Derivation, Generalization, and Applications to Cryo-EM

Amit Singer

Department of Mathematics and Program in Applied and Computational Mathematics

October 18, 2021

Acta Crystallographica Section A: Foundations and Advances, 77 (5), pp. 472–479 (2021)

Power spectrum of proteins

- The power spectrum of proteins is modelled by the Guinier law at low frequencies and the Wilson statistics at high frequencies.
- Guinier Law (1937):
  At low frequencies the power spectrum decays exponentially

$$|\hat{\phi}(\xi)|^2 \approx |\hat{\phi}(0)|^2 e^{-4\pi^2 \xi^T \frac{\Lambda}{\hat{\phi}(0)} \xi}, \quad \text{for } |\xi| \ll \frac{1}{R},$$

  where R is the size of the molecule and Λ is the moment of inertia

$$\Lambda_{ij} = \int \phi(x) x_i x_j \, dx, \quad i,j = 1,2,3.$$

- Wilson statistics (1942):
  At high frequencies the power spectrum is approximately flat

$$|\hat{\phi}(\xi)|^2 \approx \text{const.}$$

- Guinier law is derived by Taylor expansion (straightforward).
- The focus of this talk is the derivation of Wilson statistics (more challenging).

Guinier plot

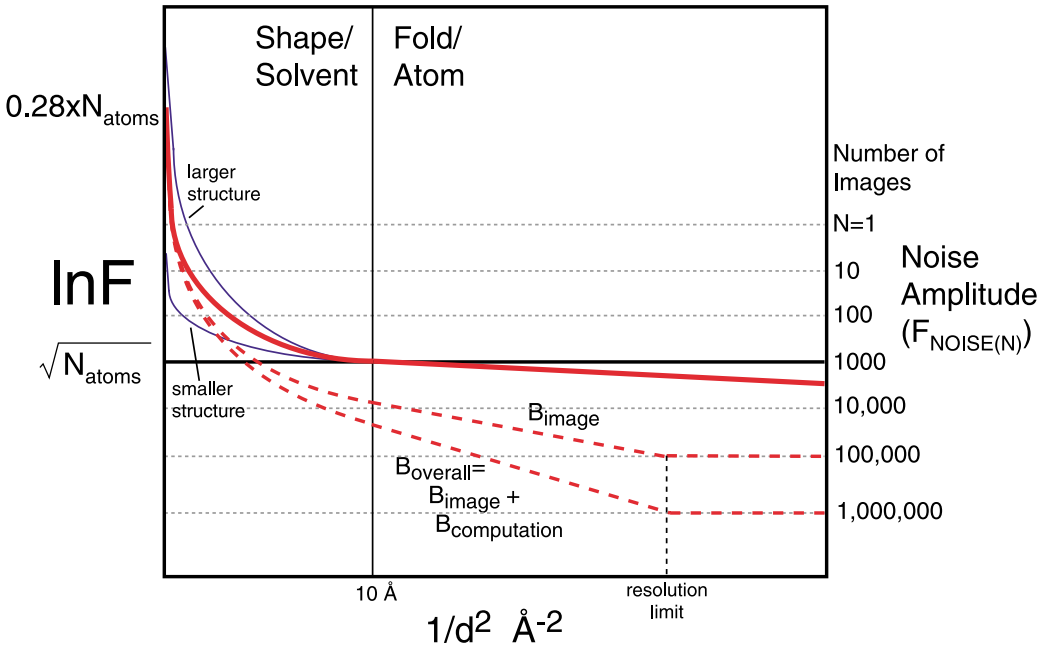


Figure 1. Schematic Guinier plot shows the natural logarithm of the spherically averaged structure factor amplitude (F) for a protein against $1/d^2$, where d is the resolution (Å). Zero angle scattering is equal to N_{atoms} carbon equivalents of the molecular mass multiplied by the solvent contrast (0.28) and places the scattering on an absolute scale. The protein scattering curve (red line) consists of a low-resolution region ($d > 10$ Å) determined by molecular shape and solvent contrast, and a high-resolution region ($d < 10$ Å) which approaches the scattering of randomly placed atoms described by Wilson statistics, which decreases only slightly with resolution and may be approximated by the horizontal line of amplitude $\sqrt{N_{\text{atoms}}}$. The high-resolution region may also have structure corresponding to fold-specific features, including α-helix and β-sheet. The average noise amplitude is $F_{\text{Noise(1)}}$ for a single image or $F_{\text{Noise(1)}}/\sqrt{N}$ after averaging N images. Low-resolution structure factor amplitudes are also shown for a large structure that might be studied by tomography and a small molecular mass particle which has a low-resolution scattering amplitude below the noise level for one image (blue lines). The experimental contrast loss for structure factors at high resolution due to imperfect images is indicated by a dotted red line labeled by its slope, the temperature factor B_{image}. Additional contrast lost due to imperfect computations gives a line with slope B_{overall}, which is the sum of temperature factors B_{image} and $B_{\text{computation}}$. The resolution limit is indicated where the structure factor curve equals the noise level, which in this example occurs at 10^6 particles for B_{overall}, but at 10^5 particles if $B_{\text{computation}} = 0$.

Rosenthal and Henderson, J. Mol. Biol. 2003

Guinier plot and B-factor sharpening

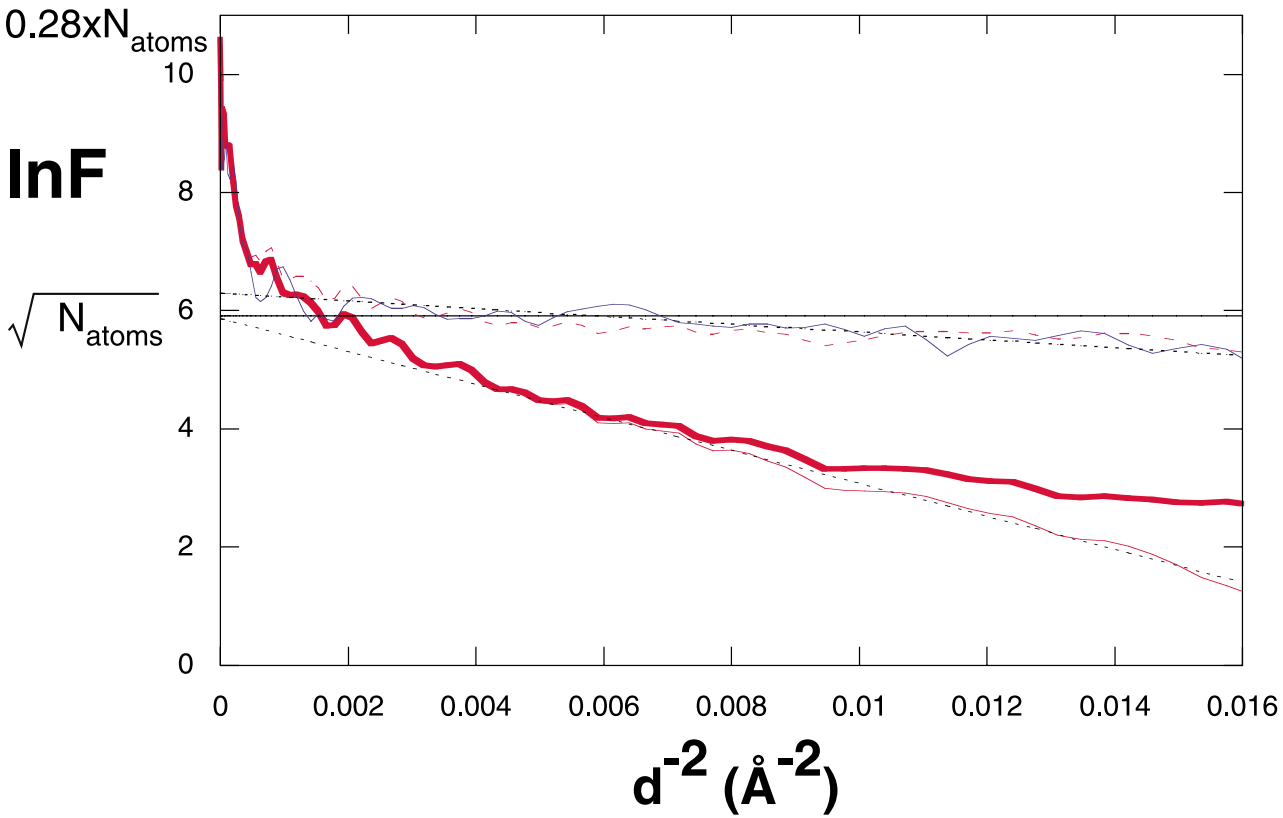


Figure 8. Guinier plot showing the natural logarithm of the spherical average of F *versus* $1/d^2$ for the experimental map (thick red line), the experimental map amplitudes weighted by C_{ref} (thin red line), weighted amplitudes after sharpening (broken red line), Wilson statistics (horizontal black line), and the X-ray model (blue line). Linear fit of data for $1/d^2$ between 0.005 and 0.015 yields $B = 1200$ Å^2 for the experimental map and slope 200 Å^2 for the model structure factors (dotted black lines). The difference, $B = -1000$ Å^2, is used to sharpen the experimental map.

Rosenthal and Henderson, J. Mol. Biol. 2003

Guinier Law

- Let $\phi(x)$ be the electrostatic potential of the molecule.
- The Fourier transform $\hat{\phi}(\xi)$ is

$$\hat{\phi}(\xi) = \int \phi(x) e^{-2\pi \imath \langle \xi, x \rangle} \, dx.$$

- Assume the molecule is centered such that its center of mass is at the origin:

$$\int \phi(x) x_i \, dx = 0, \quad i = 1, 2, 3.$$

  Equivalently, in vector notation $\int \phi(x) x \, dx = 0$.
- Assume that the molecule has size R, e.g., the molecule is compactly supported inside a ball of radius R: $\phi(x) = 0$ for $|x| > R$.
- 2nd order Taylor approximation of $e^{-2\pi \imath \langle \xi, x \rangle}$ for $|\xi| \ll \frac{1}{R}$ (low frequencies)

$$e^{-2\pi \imath \langle \xi, x \rangle} \approx 1 - 2\pi \imath \langle \xi, x \rangle + \frac{(2\pi \imath)^2}{2} \langle \xi, x \rangle^2 = 1 - 2\pi \imath \xi^T x - 2\pi^2 \xi^T x x^T \xi.$$

- The Fourier transform for $|\xi| \ll \frac{1}{R}$:

$$\hat{\phi}(\xi) \approx \int \phi(x) \left[1 - 2\pi \imath \xi^T x - 2\pi^2 \xi^T x x^T \xi \right] \, dx = \hat{\phi}(0) - 2\pi^2 \xi^T \Lambda \xi^T,$$

  where $\Lambda = \int \phi(x) x x^T \, dx$ is the inertia tensor.
  The first order term vanishes due to the center of mass assumption.

Wilson statistics: Random bag of atoms

- Suppose a "random" protein consisting of N atoms whose locations $X_1, X_2, \ldots, X_N$ are i.i.d.
- For example, each X_i could be uniformly distributed inside a container $\Omega \subset \mathbb{R}^3$, e.g., a ball or a cube, though other shapes and distributions are possible.
- The electrostatic potential is modelled as

$$\phi(x) = \sum_{i=1}^{N} f(x - X_i)$$

  where f is a bump function such as a Gaussian, or a delta function in the limit of an ideal point mass $f(x - X_i) = \delta(x - X_i)$.
- For simplicity of exposition, we assume that the atoms are identical. Otherwise, one can use a different f for the scattering from each atom type.
- The Fourier transform is given by

$$\hat{\phi}(\xi) = \sum_{i=1}^{N} \hat{f}(\xi) e^{-2\pi \imath \langle \xi, X_i \rangle} = \hat{f}(\xi) \sum_{i=1}^{N} e^{-2\pi \imath \langle \xi, X_i \rangle}$$

Wilson Statistics

$$\hat{\phi}(\xi) = \hat{f}(\xi) \sum_{i=1}^{N} e^{-2\pi\imath\langle \xi, X_i\rangle}$$

- Wilson's original argument (Nature, 1942):

$$\begin{aligned} |\hat{\phi}(\xi)|^2 &= |\hat{f}(\xi)|^2 \left| \sum_{i=1}^{N} e^{-2\pi\imath\langle \xi, X_i\rangle} \right|^2 = |\hat{f}(\xi)|^2 \left(\sum_{i,j=1}^{N} e^{-2\pi\imath\langle \xi, (X_i - X_j)\rangle} \right) \\ &= |\hat{f}(\xi)|^2 \left(N + \sum_{i \neq j} e^{-2\pi\imath\langle \xi, (X_i - X_j)\rangle} \right) \\ &\approx N|\hat{f}(\xi)|^2. \end{aligned}$$

- Since $\hat{f}(\xi) = 1$ for an ideal point mass, the power spectrum is flat: $|\hat{\phi}(\xi)|^2 \approx N$.
- Wilson argued that the sum of the complex exponentials is negligible as the terms wildly oscillate and cancel each other, especially for high frequency ξ.
- We now try to make this hand wavy argument more rigorous.
  (We are not aware of a mathematical derivation of Wilson statistics in the literature.)

Wilson Statistics

$$|\hat{\phi}(\xi)|^2 = |\hat{f}(\xi)|^2 \left(N + \sum_{i \neq j} e^{-2\pi \imath \langle \xi, (X_i - X_j) \rangle} \right).$$

- The challenge is to show that there is so much cancellation that adding $O(N^2)$ oscillating terms of size $O(1)$ is negligible compared to N.
- For a random walk, the sum of $O(N^2)$ i.i.d zero-mean random variables of variance $O(1)$ is $O(N)$ (the square-root of the number of terms).
- We want to show that the sum is negligible compared to N, so more cancellation must be happening.
- We also need to examine the role that ξ plays: Omitting $\hat{f}(\xi)$, for $\xi = 0$, clearly $|\hat{\phi}(0)|^2 = N^2$. What is the mechanism by which $|\hat{\phi}(\xi)|^2$ decays from N^2 to N as ξ increases?

Scaling argument

$$\hat{\phi}(\xi) = \hat{f}(\xi) \sum_{i=1}^{N} e^{-2\pi \imath \langle \xi, X_i \rangle}$$

- Since $X_1, \ldots, X_N$ are i.i.d, we may be tempted to apply the Central Limit Theorem (CLT) and conclude that $\hat{\phi}(\xi)$ is approximately normally distributed, with mean and variance that can be calculated.
- However, one should proceed with caution, because if the container Ω is fixed, then in the limit $N \to \infty$, the density of the atoms grows indefinitely, whereas the density of atoms in a protein is clearly bounded.
- To keep the density of atoms fixed, we allow the container Ω to grow with N and denote it Ω_N.
- Specifically, the volume of the container Ω_N should be proportional to N.
- The length-scale is therefore proportional to $N^{1/3}$, that is, $\Omega_N = N^{1/3}\Omega_1$, or $X_i = N^{1/3} Y_i$ with $Y_i \sim U(\Omega_1)$ in the uniform case, and i.i.d in general. Now,

$$\hat{\phi}(\xi) = \hat{f}(\xi) \sum_{i=1}^{N} e^{-2\pi \imath \langle \xi, N^{1/3} Y_i \rangle}$$

- The CLT cannot be applied anymore, as the random variables depend on N.

Shape of container

$$\hat{\phi}(\xi) = \hat{f}(\xi) \sum_{i=1}^{N} e^{-2\pi \imath \langle \xi, N^{1/3} Y_i \rangle}$$

- The expected value of $\hat{\phi}(\xi)$:

  $$\mathbb{E}[\hat{\phi}(\xi)] = N\hat{f}(\xi)\mathbb{E}[e^{-2\pi \imath \langle \xi, N^{1/3} Y \rangle}] = N\hat{f}(\xi) \int e^{-2\pi \imath \langle \xi, N^{1/3} y \rangle} g(y)\, dy = N\hat{f}(\xi)\hat{g}(N^{1/3}\xi),$$

  where $g(y)$ is the probability density function of Y, and $\hat{g}$ is its Fourier transform.

- The dependency on $\hat{g}(N^{1/3}\xi)$ and N being a large parameter together suggest that the decay rate of $\hat{g}$ at high frequencies is critical for analyzing Wilson statistics.

- Different container shapes and choices of g can lead to different behavior of its Fourier transform $\hat{g}$.

- Before stating known theoretical results, it is instructive to consider a couple of examples.

Uniform distribution in a ball

$$\mathbb{E}[\hat{\phi}(\xi)] = N\hat{f}(\xi)\hat{g}(N^{1/3}\xi)$$

- Here Ω_1 is a ball of radius 1, denoted B.
- The uniform density is $g_B(x) = \frac{1}{4\pi/3}\chi_B(x)$, where χ_B is the characteristic function of the ball.
- It is a radial function, a property that can readily be used to calculate its Fourier transform as
$$\hat{g}_B(\xi) = -\frac{3\cos(2\pi|\xi|)}{4\pi^2|\xi|^2} + \frac{3\sin(2\pi|\xi|)}{8\pi^3|\xi|^3}.$$
- In particular, $|\hat{g}_B(\xi)| \leq \dfrac{C}{|\xi|^2}$ for some constant $C > 0$.

Uniform distribution in a cube

- Here $\Omega_1 = [-\frac{1}{2}, \frac{1}{2}]^3$ is the unit cube.
- The uniform distribution g_C is a product of three rectangular window functions whose Fourier transform is the sinc function:

$$\hat{g}_C(\xi) = \prod_{i=1}^{3} \mathrm{sinc}(\xi_i) = \prod_{i=1}^{3} \frac{\sin(\pi\xi_i)}{\pi\xi_i}.$$

- Taking ξ along one of the axes, e.g., $\xi = (|\xi|, 0, 0)$ gives $\hat{g}_C(|\xi|, 0, 0) = \dfrac{\sin(\pi|\xi|)}{\pi|\xi|}$.
- In this case, $|\hat{g}_C(\xi)| \leq \dfrac{C}{|\xi|}$ for some $C > 0$.
- Notice that the decay of $\hat{g}_C$ in directions not normal to its faces is faster.
- For example, for $\xi = \frac{1}{\sqrt{3}}(|\xi|, |\xi|, |\xi|)$ we have

$$\left|\hat{g}_C\left(\frac{1}{\sqrt{3}}(|\xi|, |\xi|, |\xi|)\right)\right| = \frac{|\sin^3(\frac{1}{\sqrt{3}}\pi|\xi|)|}{(\frac{1}{\sqrt{3}}\pi|\xi|)^3} \leq \frac{C}{|\xi|^3}.$$

- We are now ready to state existing theoretical results about the decay rate of the Fourier transform for containers of general shape.

Decay rate of Fourier transform - Theorem 1

(see Stein and Shakarchi, Functional analysis: introduction to further topics in analysis, vol. 4., Princeton University Press, 2011 – page 336)

Theorem

1. *Suppose $\Omega \subset \mathbb{R}^d$ is a bounded region whose boundary $M = \partial\Omega$ has non-vanishing Gauss curvature at each point, then*

$$|\hat{\chi}_\Omega(\xi)| = O(|\xi|^{-\frac{d+1}{2}}), \quad \text{as } |\xi| \to \infty.$$

2. *If M has m non-vanishing principal curvatures at each point, then*

$$|\hat{\chi}_\Omega(\xi)| = O(|\xi|^{-(m+2)/2}), \quad \text{as } |\xi| \to \infty.$$

The decay rates previously observed for the three-dimensional ball ($d = 3$ or $m = 2$) and the cube ($m = 0$) are particular cases of Theorem 1.

Decay rate of spherically averaged Fourier transform

Although the decay rate in different directions could be different (as the example of the cube illustrates), for a large family of containers (convex sets and open sets with sufficiently smooth boundary surface), Theorem 2 asserts that the spherical average of the power spectrum has the same decay rate as that of the ball.

L. Brandolini, S. Hofmann, and A. Iosevich, "Sharp rate of average decay of the Fourier transform of a bounded set," Geometric & Functional Analysis GAFA, vol. 13, no. 4, pp. 671–680, 2003.

Theorem

Suppose $\Omega \subset \mathbb{R}^d$ is a convex body or an open bounded set whose boundary $\partial\Omega$ is $C^{3/2}$. Then,

$$\int_{S^{d-1}} |\hat{\chi}_\Omega(k\omega)|^2 \, d\omega = O(k^{-(d+1)}), \quad \text{as } k \to \infty.$$

Here $k = |\xi|$ is the radial frequency and S^{d-1} is the unit sphere in $\mathbb{R}^d$.

We are now in position to state and prove our main result that fully characterizes the regime of validity of Wilson statistics.

Validity regime of Wilson statistics

Theorem (S, 2021)

1. *For the random bag of atoms model, the expected power spectrum is given by*

$$\mathbb{E}\left[|\hat{\phi}(\xi)|^2\right] = |\hat{f}(\xi)|^2 \left(N + N(N-1)\left|\hat{g}(N^{1/3}\xi)\right|^2\right).$$

2. *If the container is a convex body or an open set with a $C^{3/2}$ boundary surface, and the atomic locations are uniformly distributed in the container, then the expected spherically-averaged power spectrum satisfies*

$$\mathbb{E}\left[\frac{1}{4\pi}\int_{S^2} |\hat{\phi}(k\omega)|^2 \, d\omega\right] = |\hat{f}(k)|^2 \left(N + o(N)\right),$$

   for $k \gg N^{-1/12}$.

3. *If the Fourier transform of the density g satisfies $|\hat{g}(\xi)| \leq C|\xi|^{-\alpha}$, then*

$$\mathbb{E}\left[|\hat{\phi}(\xi)|^2\right] = |\hat{f}(\xi)|^2 \left(N + o(N)\right), \quad \text{for } |\xi| \gg N^{-\beta},$$

   where $\beta = \dfrac{2\alpha - 3}{6\alpha}$.

Proof - Part I

- Start with Wilson's original approach:

$$|\hat{\phi}(\xi)|^2 = \left|\hat{f}(\xi)\sum_{i=1}^{N} e^{-2\pi\imath\langle\xi,N^{1/3}Y_i\rangle}\right|^2 = |\hat{f}(\xi)|^2 \sum_{i,j=1}^{N} e^{-2\pi\imath\langle\xi,N^{1/3}(Y_i-Y_j)\rangle}$$

$$= |\hat{f}(\xi)|^2 \left(N + \sum_{i\neq j} e^{-2\pi\imath\langle\xi,N^{1/3}(Y_i-Y_j)\rangle}\right).$$

- Since the Y_i's are i.i.d, the expected power spectrum satisfies

$$\mathbb{E}\left[|\hat{\phi}(\xi)|^2\right] = |\hat{f}(\xi)|^2 \left(N + \sum_{i\neq j} \mathbb{E}\left[e^{-2\pi\imath\langle\xi,N^{1/3}(Y_i-Y_j)\rangle}\right]\right)$$

$$= |\hat{f}(\xi)|^2 \left(N + \sum_{i\neq j} \mathbb{E}\left[e^{-2\pi\imath\langle\xi,N^{1/3}Y_i\rangle}\right] \overline{\mathbb{E}\left[e^{2\pi\imath\langle\xi,N^{1/3}Y_j\rangle}\right]}\right)$$

$$= |\hat{f}(\xi)|^2 \left(N + N(N-1)\left|\mathbb{E}\left[e^{-2\pi\imath\langle\xi,N^{1/3}Y\rangle}\right]\right|^2\right)$$

$$= |\hat{f}(\xi)|^2 \left(N + N(N-1)\left|\hat{g}(N^{1/3}\xi)\right|^2\right).$$

Proof - Part II

- We proved

$$\mathbb{E}\left[|\hat{\phi}(\xi)|^2\right] = |\hat{f}(\xi)|^2 \left(N + N(N-1)\left|\hat{g}(N^{1/3}\xi)\right|^2\right).$$

- Assuming f (hence also $\hat{f}$) are radial functions, the expectation of the spherically-averaged power spectrum satisfies

$$\mathbb{E}\left[\frac{1}{4\pi}\int_{S^2}|\hat{\phi}(k\omega)|^2\,d\omega\right] = |\hat{f}(k)|^2\left(N + N(N-1)\frac{1}{4\pi}\int_{S^2}\left|\hat{g}(N^{1/3}k\omega)\right|^2\,d\omega\right).$$

- Theorem 2 with $d = 3$ implies

$$N(N-1)\frac{1}{4\pi}\int_{S^2}\left|\hat{g}(N^{1/3}k\omega)\right|^2\,d\omega = O(N^{2/3}k^{-4}).$$

- This term is negligible compared to N for $k \gg N^{-1/12}$, proving part II:

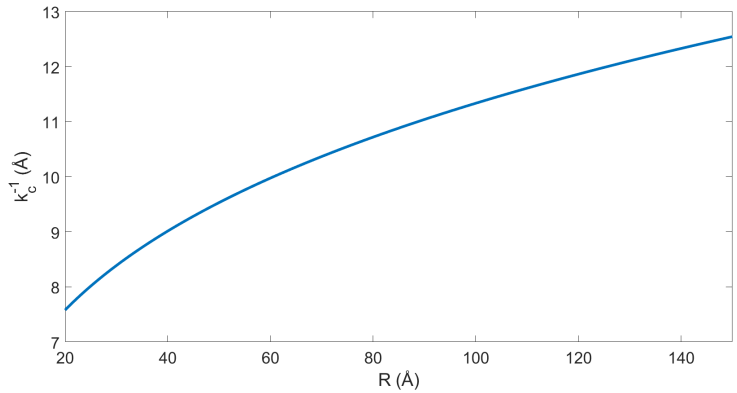
$$\mathbb{E}\left[\frac{1}{4\pi}\int_{S^2}|\hat{\phi}(k\omega)|^2\,d\omega\right] = |\hat{f}(k)|^2\left(N + o(N)\right),$$

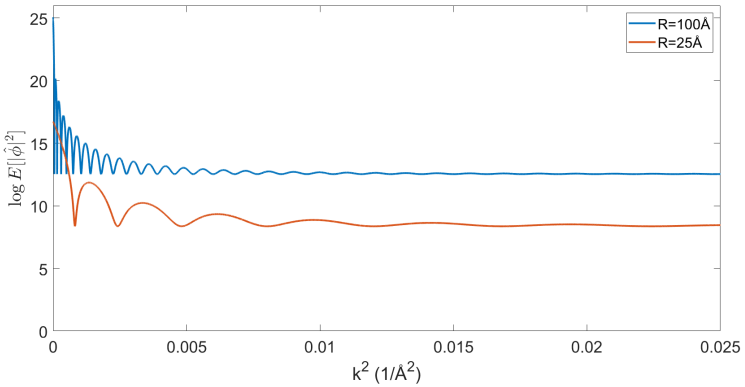
for $k \gg N^{-1/12}$.

Theoretical Guinier plots and cutoff frequencies

- The protein density is approximately $\rho \approx 0.8\,\mathrm{Da}/\mathrm{\AA}^3$. The cutoff frequency is

$$k_c = 0.28 R^{-1/4} = 0.31 M_W^{-1/12}$$





- For the larger molecule with $R = 100$Å the power spectrum is approximately flat above $k^2 = 0.01\mathrm{\AA}^{-2}$ corresponding to 10Å resolution, whereas for the smaller molecule with $R = 25$Å the transition occurs closer to $k^2 = 0.015\mathrm{\AA}^{-2}$, or 8.2Å resolution.
- The cutoff frequencies are in agreement with empirical evidence about the validity regime of Wilson statistics.

Transition to Wilson statistics

$$\mathbb{E}\left[|\hat{\phi}(\xi)|^2\right] = |\hat{f}(\xi)|^2 \left(N + N(N-1)\left|\hat{g}(N^{1/3}\xi)\right|^2\right).$$

$$N(N-1)\frac{1}{4\pi}\int_{S^2}\left|\hat{g}(N^{1/3}k\omega)\right|^2 d\omega = O(N^{2/3}k^{-4}).$$

- The spherically-averaged power spectrum decays to its high frequency limit as k^{-4}.
- At first, the $1/12$ exponent of the cutoff frequency $k_0 = N^{-1/12}$ might seem mysterious.
- In hindsight, it is simply the product of the dimension $d = 3$ that resulted in the scaling of $N^{1/3}$ and the decay rate exponent of k^{-4}.

Statistical fluctuations

- Even though we characterized the expected power spectrum, one may wonder whether the statistical fluctuations of the power spectrum could overwhelm its mean.

- This turns out not to be the case.

- Similar to the derivation of Wilson statistics, one can show that if $|\hat{g}(\xi)| \leq C|\xi|^{-2}$ then
$$\mathbb{E}\left[|\hat{\phi}(\xi)|^4\right] = N^2|\hat{f}(\xi)|^4 + o(N^2), \quad \text{for } |\xi| \gg N^{-1/12}.$$

- Since $\mathbb{E}\left[|\hat{\phi}(\xi)|^2\right] = N|\hat{f}(\xi)|^2 + o(N)$ for $|\xi| \gg N^{-1/12}$, it follows that for $|\xi| \gg N^{-1/12}$
$$\mathrm{Var}(|\hat{\phi}(\xi)|^2) = \mathbb{E}\left[|\hat{\phi}(\xi)|^4\right] - \mathbb{E}\left[|\hat{\phi}(\xi)|^2\right]^2 = o(N^2).$$

- In other words, the standard deviation of the power spectrum is $o(N)$, so the fluctuation is smaller than the mean value.

Generalizations and applications to cryo-EM

Existing applications:

- Map sharpening, B-factor correction, B-factor flattening, or B-factor sharpening: estimate the B-factor from the Guinier plot, and boost medium and high frequency components to agree with Wilson statistics.
- B-factor sharpening increases the contrast of many structural features of the map and helps to model the atomic structure.
- Wilson statistics is also used to reason about and extrapolate the number of particles required to high resolution reconstruction.

Application to Bayesian Inference 3-D Refinement

- A potential application of Wilson statistics is to 3-D refinement.
- The Bayesian inference framework underlying RELION requires the covariance matrix of $\hat{\phi}$ and approximates it with a diagonal matrix.
- For tractable computation, the variance is further assumed to be a radial function.

Scheres, JSB 2012; Scheres, JMB 2012

- The covariance matrix can be determined for the random bag of atoms model underlying Wilson statistics:

$$\mathrm{Cov}[\hat{\phi}](\xi_1,\xi_2) = \mathbb{E}[\hat{\phi}(\xi_1)\overline{\hat{\phi}(\xi_2)}] - \mathbb{E}[\hat{\phi}(\xi_1)]\mathbb{E}[\overline{\hat{\phi}(\xi_2)}]$$

- The two terms are given by

$$\mathbb{E}[\hat{\phi}(\xi_1)\overline{\hat{\phi}(\xi_2)}] = \hat{f}(\xi_1)\overline{\hat{f}(\xi_2)}\left[N\hat{g}(N^{1/3}(\xi_1-\xi_2)) + N(N-1)\hat{g}(N^{1/3}\xi_1)\overline{\hat{g}(N^{1/3}\xi_2)}\right],$$

and

$$\mathbb{E}[\hat{\phi}(\xi_1)]\mathbb{E}[\overline{\hat{\phi}(\xi_2)}] = N^2\hat{f}(\xi_1)\overline{\hat{f}(\xi_2)}\hat{g}(N^{1/3}\xi_1)\overline{\hat{g}(N^{1/3}\xi_2)}.$$

- Therefore,

$$\mathrm{Cov}[\hat{\phi}](\xi_1,\xi_2) = N\hat{f}(\xi_1)\overline{\hat{f}(\xi_2)}\left[\hat{g}(N^{1/3}(\xi_1-\xi_2)) - \hat{g}(N^{1/3}\xi_1)\overline{\hat{g}(N^{1/3}\xi_2)}\right].$$

Application to Bayesian Inference 3-D Refinement

$$\mathrm{Cov}[\hat{\phi}](\xi_1,\xi_2) = N\hat{f}(\xi_1)\overline{\hat{f}(\xi_2)}\left[\hat{g}(N^{1/3}(\xi_1-\xi_2)) - \hat{g}(N^{1/3}\xi_1)\overline{\hat{g}(N^{1/3}\xi_2)}\right].$$

- This result implies a vast reduction in the number of parameters needed to describe the covariance matrix.
- In general, for a 3-D map represented as an array of L^3 voxels, the covariance matrix is of size $L^3 \times L^3$ which requires $O(L^6)$ entries, which is prohibitively large.
- The result suggests that the covariance depends on only $O(L^3)$ parameters.
- Furthermore, approximating $\hat{g}(\xi)$ by a radial function implies that the covariance depends on just $O(L)$ parameters, the same number of parameters in the existing Bayesian inference method for 3-D iterative refinement.

Application to Bayesian Inference 3-D Refinement

$$\mathrm{Cov}[\hat{\phi}](\xi_1,\xi_2) = N\hat{f}(\xi_1)\overline{\hat{f}(\xi_2)}\left[\hat{g}(N^{1/3}(\xi_1-\xi_2)) - \hat{g}(N^{1/3}\xi_1)\overline{\hat{g}(N^{1/3}\xi_2)}\right].$$

- Comparing the two terms in the covariance, the decay of $\hat{g}$ implies that $|\hat{g}(N^{1/3}(\xi_1-\xi_2))| \gg |\hat{g}(N^{1/3}\xi_1)\overline{\hat{g}(N^{1/3}\xi_2)}|$ whenever $|\xi_1|, |\xi_2| \gg N^{-1/3}$.
- Therefore, for $|\xi_1|, |\xi_2| \gg N^{-1/3}$

$$\mathrm{Cov}[\hat{\phi}](\xi_1,\xi_2) = N\hat{f}(\xi_1)\overline{\hat{f}(\xi_2)}\hat{g}(N^{1/3}(\xi_1-\xi_2))(1+o(1)).$$

- $\hat{g}(N^{1/3}(\xi_1-\xi_2))$ is largest for $\xi_1 = \xi_2$ and decays with increasing distance $|\xi_1-\xi_2|$.
- The covariance matrix is approximately a band matrix with bandwidth $O(N^{-1/3})$.
- $N^{-1/3}$ is a very low frequency corresponding to resolution of the size of the protein (as implied by the $N^{1/3}$ scaling). Therefore, the covariance is well approximated by a band matrix with a small number of diagonals.
- This serves as a theoretical justification for the diagonal approximation in the Bayesian inference framework, as correlations of Fourier coefficients with $|\xi_1-\xi_2| \gg N^{-1/3}$ are negligible.
- On the flip side, correlations for which $|\xi_1-\xi_2| \ll N^{-1/3}$ should not be ignored and correctly accounting for them could potentially lead to further improvement.

Application to Bayesian Inference 3-D Refinement

$$\mathrm{Cov}[\hat{\phi}](\xi_1,\xi_2) = N\hat{f}(\xi_1)\overline{\hat{f}(\xi_2)}\left[\hat{g}(N^{1/3}(\xi_1-\xi_2)) - \hat{g}(N^{1/3}\xi_1)\overline{\hat{g}(N^{1/3}\xi_2)}\right].$$

- Note that the diagonal of the covariance matrix satisfies

$$\mathrm{Var}(\hat{\phi}(\xi)) = \mathrm{Cov}[\hat{\phi}](\xi,\xi) = N|\hat{f}(\xi)|^2\left[1 - |\hat{g}(N^{1/3}\xi)|^2\right].$$

- The variance vanishes for $\xi = 0$ because $\hat{\phi}(0) = N$ regardless of atomic positions.
- In existing Bayesian inference approaches, the mean of each frequency voxel is assumed to be zero.
- However, for the random bag of atoms model the variance dominates the squared mean only for $|\xi| \gg N^{-1/12}$, which is the validity regime of Wilson statistics.
- It is justified to assume a zero-mean signal only for high frequencies, but not at low frequencies.
- Including an explicit (approximately radial) non-zero mean in the Bayesian inference framework may therefore bring further improvement.

Method of moments for 3-D reconstruction

- Kam (1980) suggested an autocorrelation method for determining the 3-D structure from the moment statistics of the noisy 2-D images.
- Kam's original method is limited to uniform distribution of viewing directions.
- Recently, Kam's approach was extended to non-uniform distributions (Sharon et al, Inverse Problems 2020) and to reconstruction directly from micrographs without particle picking (Bendory et al, Inverse Problems 2019).
- These methods are based on sample moments and higher order spectra, but so far have been limited to relatively low resolution.

Wilson statistics and the autocorrelation approach

$$\mathbb{E}\left[|\hat{\phi}(\xi)|^2\right] \quad = \quad |\hat{f}(\xi)|^2 \left(N + N(N-1) \left|\hat{g}(N^{1/3}\xi)\right|^2 \right)$$

- Wilson statistics highlights two difficulties for the moment-based approach for reconstruction directly from micrographs without particle picking.
- First, the power spectrum drops from N^2 at low frequencies to N at medium-high frequencies. The method would give much more emphasis to fitting the low-frequency content of the map.
- Second, and perhaps more discouraging, there are many different molecules (corresponding to different realizations of random placement of atoms) that have approximately the same flat power spectrum at high frequencies. Recovering the 3-D map from its power spectrum is therefore ill-conditioned, especially at medium-high frequencies due to the "universality" of the power spectrum.
- What about higher order spectra?

Higher order Wilson statistics

- Next, calculate the expected bispectrum for the random "bag of atoms" model in order to find its implication on the third order autocorrelation.
- Use $\hat{\phi}(\xi) = \hat{f}(\xi) \sum_{i=1}^{N} e^{-2\pi\imath\langle \xi, X_i\rangle}$ to calculate the expectation of $\hat{\phi}(\xi_1)\hat{\phi}(\xi_2)\hat{\phi}(\xi_3)$:

$$
\begin{aligned}
&\mathbb{E}\left[\hat{\phi}(\xi_1)\hat{\phi}(\xi_2)\hat{\phi}(\xi_3)\right] \\
&= \hat{f}(\xi_1)\hat{f}(\xi_2)\hat{f}(\xi_3)\mathbb{E}\left[\sum_{i,j,k=1}^{N} e^{-2\pi\imath\langle \xi_1, N^{1/3}Y_i\rangle} e^{-2\pi\imath\langle \xi_2, N^{1/3}Y_j\rangle} e^{-2\pi\imath\langle \xi_3, N^{1/3}Y_k\rangle}\right] \\
&= \hat{f}(\xi_1)\hat{f}(\xi_2)\hat{f}(\xi_3)\left[N\hat{g}\left(N^{1/3}(\xi_1+\xi_2+\xi_3)\right)\right. \\
&+N(N-1)\hat{g}\left(N^{1/3}(\xi_1+\xi_2)\right)\hat{g}\left(N^{1/3}\xi_3\right) \\
&+N(N-1)\hat{g}\left(N^{1/3}(\xi_1+\xi_3)\right)\hat{g}\left(N^{1/3}\xi_2\right) \\
&+N(N-1)\hat{g}\left(N^{1/3}(\xi_2+\xi_3)\right)\hat{g}\left(N^{1/3}\xi_1\right) \\
&\left.+\ N(N-1)(N-2)\hat{g}\left(N^{1/3}\xi_1\right)\hat{g}\left(N^{1/3}\xi_2\right)\hat{g}\left(N^{1/3}\xi_3\right)\right]
\end{aligned}
$$

Higher order Wilson statistics

- Similar to the power spectrum $|\hat{\phi}(\xi)|^2$ which is the Fourier transform of the autocorrelation function, the bispectrum $\hat{\phi}(\xi_1)\hat{\phi}(\xi_2)\hat{\phi}(-(\xi_1+\xi_2))$ is the Fourier transform of the triple-correlation function.
- The bispectrum, like the power spectrum, is also shift-invariant. As such, it plays an important role in various autocorrelation analysis techniques.
- For the expected bispectrum we set $\xi_1+\xi_2+\xi_3=0$:

$$\begin{aligned}\mathbb{E}\left[\hat{\phi}(\xi_1)\hat{\phi}(\xi_2)\hat{\phi}(\xi_3)\right] &= \hat{f}(\xi_1)\hat{f}(\xi_2)\hat{f}(\xi_3)\Big[N+N(N-1)\left|\hat{g}\left(N^{1/3}\xi_1\right)\right|^2 \\ &\quad +N(N-1)\left|\hat{g}\left(N^{1/3}\xi_2\right)\right|^2+N(N-1)\left|\hat{g}\left(N^{1/3}\xi_3\right)\right|^2 \\ &\quad +N(N-1)(N-2)\hat{g}\left(N^{1/3}\xi_1\right)\hat{g}\left(N^{1/3}\xi_2\right)\hat{g}\left(N^{1/3}\xi_3\right)\Big]\end{aligned}$$

- The bispectrum drops from N^3 for $\xi_1=\xi_2=\xi_3=0$ to N at high frequencies.
- This drop is even more pronounced than that of the power spectrum that decreases from N^2 to N.
- This may lead to numerical difficulties in inverting the bispectrum as it has a large dynamic range, e.g., it spans eight orders of magnitude for $N=10^4$.

Discussion

- Wilson statistics is an instance of a universality phenomenon: all proteins regardless of their shape and specific atomic positions exhibit a similar power spectrum at high frequencies. This universality is a blessing and a curse at the same time. On the one hand, it enables to correct the magnitudes of the Fourier coefficients of the reconstructed map so they agree with the theoretical prediction. On the other hand, it implies that the high frequency part of the power spectrum is not particularly useful for structure determination, as it does not discriminate between molecules.

- The generalization of Wilson statistics to higher order spectra shows that the bispectrum is ill-conditioned and becomes flat at high frequencies. These observations may help explain difficulties of the autocorrelation approach as a high resolution reconstruction method.

- Wilson statistics ignores correlations between atomic positions in the protein. It is well known that the power spectrum deviates from Wilson statistics at frequencies that correspond to interatomic distances associated with secondary structure such as α-helices which produce a peak at 10Å and beta-sheets which produce a peak at 4.5Å.

Summary

- Presented the first formal mathematical derivation of Wilson statistics.
- $k_0 = N^{-1/12}$ is the (dimensionless) frequency above which the power spectrum is approximately flat.
- The random bag of atoms model enables the derivation of useful statistics beyond the power spectrum, such as the 3-D covariance and higher order spectra (bispectrum).
- These generalizations of Wilson statistics can potentially be applied to other aspects of the computational pipeline of single-particle analysis beyond B-factor correction.

Thank You!


NIH
NIGMS


NSF


AIR FORCE OFFICE OF SCIENTIFIC RESEARCH
UNITED STATES AIR FORCE


GORDON AND BETTY
MOORE
FOUNDATION


SIMONS
FOUNDATION