5. X–ray diffraction and biomolecular structure

The first detailed experimental information about the structure of biological molecules came from X–ray diffraction measurements. We recall that if a particle scatters from a sample, shifting its energy by $\hbar\omega$ and its momentum by $\hbar\vec{\mathbf{q}}$, then the amplitude for this scattering event must be proportional to the $(\vec{\mathbf{q}},\omega)$ spatiotemporal Fourier component of the relevant density in the sample. For an electromagnetic wave what matters is (roughly) the charge density. Thus, the cross–section for elastic ($\omega = 0$) scattering is

$$\sigma(\vec{\mathbf{q}}) \propto \left| \int d^3x\, e^{i\vec{\mathbf{q}}\cdot\vec{\mathbf{x}}} \rho(\vec{\mathbf{x}}) \right|^2 . \qquad \text{(A247)}$$

It is useful to have in mind the geometry [ref to a Fig!]. If the X–ray photons approach the sample collimated along the $\hat{x}$ axis, they have an initial wavevector $\vec{\mathbf{k}}_0 = k\hat{x}$, where as usual $k = 2\pi/\lambda$, with λ the wavelength. If they emerge with a final wavevector $\vec{\mathbf{k}}_f$ at an angle θ relative to the $\hat{x}$ axis, then $\vec{\mathbf{q}} \equiv \vec{\mathbf{k}}_f - \vec{\mathbf{k}}_0$, and the magnitude of the scattering vector (or, up to a factor $\hbar$, momentum transfer) is

$$\begin{aligned} |\vec{\mathbf{q}}| &= |\vec{\mathbf{k}}_f - \vec{\mathbf{k}}_0| && \text{(A248)} \\ &= \sqrt{|\vec{\mathbf{k}}_f - \vec{\mathbf{k}}_0|^2} && \text{(A249)} \\ &= \sqrt{|\vec{\mathbf{k}}_f|^2 - 2\vec{\mathbf{k}}_f\cdot\vec{\mathbf{k}}_0 + |\vec{\mathbf{k}}_0|^2} && \text{(A250)} \\ &= \sqrt{k^2 - 2k^2\cos\theta + k^2} && \text{(A251)} \\ &= \sqrt{2k^2(1-\cos\theta)} = 2k\sin(\theta/2). && \text{(A252)} \end{aligned}$$

Thus scattering by a small angle corresponds to a small momentum transfer. The classic results about X–ray diffraction concern the case where the density profile is periodic, as in a crystal. If the periodicity corresponds to displacement by d (let's think along one dimension, for the moment), then the density can be expressed as a discrete Fourier series, which means [from Eq (A247)] that $\sigma(\vec{\mathbf{q}})$ will have delta functions at $|\vec{\mathbf{q}}| = 2\pi n/d$, with n an integer. Combining this with Eq (A252), we find the angles which satisfy the "Bragg condition,"

$$2\pi n/d = (4\pi/\lambda)\sin(\theta/2) \Rightarrow \sin(\theta/2) = n\lambda/2d. \qquad \text{(A253)}$$

[I think this is a bit off the usual way of stating the condition (2's in the wrong places); check!]

The first great triumph of X–ray diffraction in elucidating the structure of biological molecules came with the structure of DNA. This is an often told, and often distorted, piece of scientific history. Watson and Crick predicted the structure of DNA by arguing that a few key facts about the molecule, when combined with the rules of chemical bonding, where enough to suggest an interesting structure that would have consequences for the mechanisms of genetic inheritance. It was known that DNA was composed of four different kinds of nucleotide bases: adenine (A), thymine (A), guanine (G) and cytosine (C). Importantly, Chargaff had surveyed the DNA of many organisms and shown that while the ratios of A to G, for example, vary enormously, the ratios A/T and C/G do not. Watson and Crick realized that the molecular structures of the bases are such that A and T can form favorable hydrogen bonds, as can C and G; further, the resulting hydrogen bonded base pairs are the same size, and thus could fit comfortably into a long polymer, as shown in Fig 171. Piling on top of one another, the base pairs would also experience a favorable "stacking" interaction among the π–bonded electrons in their rings. Finally, if one looks carefully at all the bond angles where the planar bases connect to the sugars and phosphate backbone, each successive base pair must rotate relative to its neighbor, and although there is some flexibility the favored angle was predicted to be $2\pi/10$ radians, or 36°.

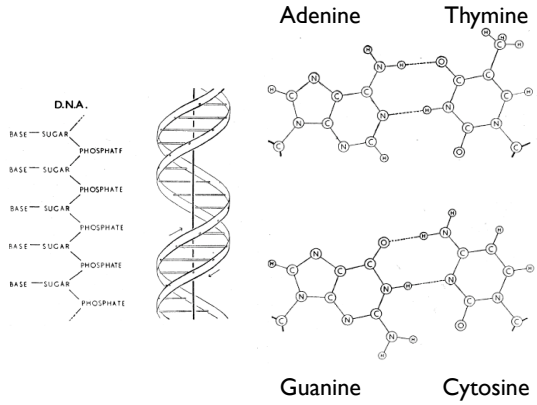

FIG. 171 The structure of DNA, from Watson and Crick (1953b). At left, the polymeric pattern of bases, sugars and phosphates, and the famous double helix. At right, the pairings A/T and G/C, illustrating the similar sizes of the correct pairs. Note that the donor/acceptor pattern of hydrogen bonds discriminates against the incorrect A/C and G/T pairings.

Quite independently of his collaboration with Watson, Crick has been interested in the structure of helical molecules, and in the X–ray diffraction patterns that they should produce. Thus, when Watson and Crick realized that the structure of DNA might be a helix, they were in a position to calculate what the diffraction patterns should look like, and thus compare with the data emerging from the work of Franklin, Wilkins and collaborators. So, let's look at the theory of diffraction from a helix.

It's best to describe a helix in cylindrical coordinates: z along the axis of the helix, r outward from its center, and an angle ϕ around the axis. Helical symmetry is the statement that translations along z are equivalent

to rotations of the angle ϕ. Thus, a continuous helical structure would have the property that

$$\rho(z, r, \phi) = \rho(z + d, r, \phi + 2\pi d/\ell), \tag{A254}$$

for any displacement d, where ℓ is the displacement corresponding to a complete rotation. For a discrete helical structure, the same equation is true, but only for values of d that are integer multiples of a fundamental spacing d_0.

For the continuous helix, the dependence on the two variables z and ϕ really collapses to a dependence on one combined variable,

$$\rho(z, r, \phi) = g(r, \phi - 2\pi z/\ell). \tag{A255}$$

We know that any function of angle can be expanded as a discrete Fourier series,

$$f(\phi) = \sum_{n=-\infty}^{\infty} \tilde{f}_n e^{-in\phi}, \tag{A256}$$

so in this case we have

$$\rho(z, r, \phi) = \sum_{n=-\infty}^{\infty} \tilde{g}_n(r) e^{-in(\phi - 2\pi z/\ell)}. \tag{A257}$$

Our task is to compute

$$\int d^3x\, e^{i\vec{\mathbf{q}}\cdot\vec{\mathbf{x}}} \rho(\vec{\mathbf{x}}). \tag{A258}$$

In cylindrical coordinates, we can write $\vec{\mathbf{q}} = (q_z\hat{z}, \vec{\mathbf{q}}_\perp)$, so that $\vec{\mathbf{q}}\cdot\vec{\mathbf{x}} = q_z z + q_\perp r\cos\phi$, where we choose the origin of the angle ϕ to make things simple and $q_\perp = |\vec{\mathbf{q}}_\perp|$. Thus we have

$$e^{i\vec{\mathbf{q}}\cdot\vec{\mathbf{x}}} = e^{iq_z z} e^{iq_\perp r\cos\phi} \tag{A259}$$

$$= e^{iq_z z} \sum_{n=-\infty}^{\infty} J_n(q_\perp r) e^{in\phi}, \tag{A260}$$

where [check the conventions for the definition of the Bessel function!]

$$J_n(u) = \int_0^{2\pi} \frac{d\phi}{2\pi} e^{-in\phi} e^{iu\cos\phi} \tag{A261}$$

are Bessel functions. Putting Eq (A260) together with the consequences of helical symmetry in Eq (A257), we have

$$\int d^3x\, e^{i\vec{\mathbf{q}}\cdot\vec{\mathbf{x}}} \rho(\vec{\mathbf{x}}) = \int_{-\infty}^{\infty} dz \int_0^{\infty} dr r \int_0^{2\pi} d\phi e^{iq_z z} \sum_{n=-\infty}^{\infty} J_n(q_\perp r) e^{in\phi} \sum_{m=-\infty}^{\infty} \tilde{g}_m(r) e^{-im(\phi - 2\pi z/\ell)} \tag{A262}$$

$$= \sum_{n,m-\infty}^{\infty} \int_{-\infty}^{\infty} dz\, e^{iq_z z} e^{-i2\pi mz/\ell} \int_0^{\infty} dr r\, J_n(q_\perp r)\tilde{g}_m(r) \int_0^{2\pi} e^{in\phi} e^{-im\phi}. \tag{A263}$$

We see that the integral over ϕ forces $m = n$, and the integral over z generates delta functions at $q_z = 2\pi n/\ell$. Thus, for a continuous helix we expect that the X–ray scattering cross section will behave as

$$\sigma(q_z, q_\perp) \propto \sum_{n-\infty}^{\infty} \delta(q_z - 2\pi n/\ell) \left| \int_0^{\infty} dr r\, J_n(q_\perp r)\tilde{g}_n(r) \right|^2. \tag{A264}$$

In particular, if most of the density sits at a distance R from the center of the helix (which is not a bad approximation for DNA, since the phosphate groups have much more electron density than the rest of the molecule), then

$$\sigma(q_z, q_\perp) \sim \sum_{n-\infty}^{\infty} \delta(q_z - 2\pi n/\ell) \left| J_n(q_\perp R) \right|^2. \tag{A265}$$

Equation (A265) is telling us that diffraction from a helix generates a series of "layer lines" at $q_z = 2\pi n/\ell$, and from their spacing we should be able to read off the "pitch" of the helix, the distance ℓ along the $\hat{z}$ axis corresponding to a complete turn. Further, if we look along a single layer line, we should see an intensity varying as

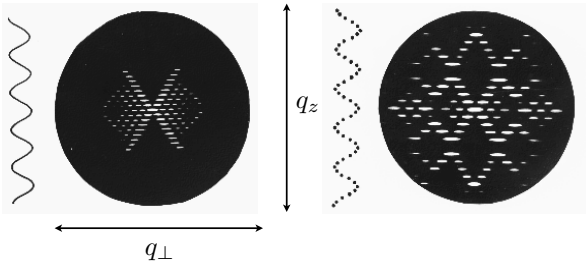

FIG. 172 Diffraction from continuous (left) and discrete (right) helices; Holmes (1998).

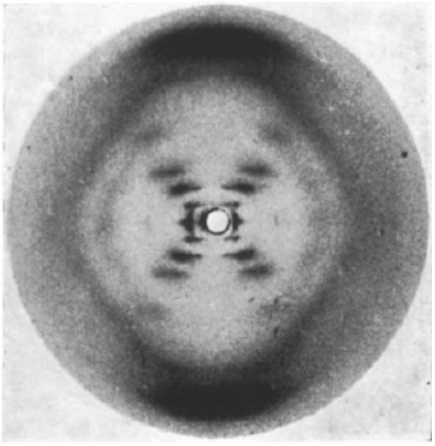

FIG. 173 The justly famous photograph 51, showing the diffraction from DNA molecules pulled into a fiber, from Franklin & Gosling (1953).

$\sim |J_n(q_\perp R)|^2$. What is important here about the Bessel functions is that for small $q_\perp$ we have $J_n(q_\perp R) \propto (q_\perp R)^n$, and the first peak of the n^{th} Bessel function occurs at a point roughly proportional to n. The resulting pattern is shown schematically in Fig 172.

Problem 180: Bessel functions. Verify the statements about Bessel functions made above, in enough detail to understand the diffraction patterns shown in Fig 172.

Let's see what happens when we move from the continuous to the discrete helix. To keep things simple, suppose that all the density indeed is concentrated at a distance R from the center of the helix, so that

$$\rho(\vec{\mathbf{x}}) = \frac{1}{R}\delta(r-R)\sum_n \delta(z-nd_0)\delta(\phi - n\phi_0), \quad \text{(A266)}$$

where the rotation from one element to the next $\phi_0 = 2\pi d_0/\ell$; notice that we don't really require ℓ/d_0 to be an integer. Now we have

$$\int d^3x\, e^{i\vec{\mathbf{q}}\cdot\vec{\mathbf{x}}}\rho(\vec{\mathbf{x}}) = \int_{-\infty}^{\infty} dz \int_0^{\infty} dr r \int_0^{2\pi} d\phi e^{iq_z z} \sum_{n=-\infty}^{\infty} J_n(q_\perp r)e^{in\phi}\frac{1}{R}\delta(r-R)\sum_{m=-\infty}^{\infty}\delta(z-md_0)\delta(\phi - m\phi_0) \quad \text{(A267)}$$

$$= \sum_{n=-\infty}^{\infty} J_n(q_\perp R)\sum_{m=-\infty}^{\infty}\int_{-\infty}^{\infty} dz\,\delta(z-md_0)e^{iq_z z} \times \int_0^{2\pi} d\phi\,\delta(\phi - m\phi_0)e^{in\phi} \quad \text{(A268)}$$

$$= \sum_{n=-\infty}^{\infty} J_n(q_\perp R)\sum_{m=-\infty}^{\infty} e^{im(n\phi_0 + q_z d_0)} \quad \text{(A269)}$$

$$= \sum_{n=-\infty}^{\infty} J_n(q_\perp R)\sum_{m=-\infty}^{\infty} \delta(n\phi_0 + q_z d_0 - 2\pi m) \quad \text{(A270)}$$

$$\propto \sum_{m=-\infty}^{\infty}\sum_{n=-\infty}^{\infty} J_n(q_\perp R)\delta(q_z + 2\pi n/\ell - 2\pi m/d_0). \quad \text{(A271)}$$

Thus the discrete helix involves a double sum of terms. If we set $m = 0$ we have the results for the continuous helix. But the sum over $m \neq 0$ causes the whole "X" pattern of the continuous helix to be repeated with centers at $(q_z = 2\pi m/d_0, q_\perp = 0)$; the line $q_\perp = 0$ is often called the meridian, and so the extra peaks centered on $(q_z = 2\pi m/d_0, q_\perp = 0)$ are called meridional refections. All of this is shown in Fig 172. Just as the spacing of the layer lines allows us to measure the helical pitch ℓ, the spacing of the meridional reflections allows us to measure the spacing d_0 between discrete elements along the helix.

At this point you know what Watson and Crick knew [maybe put in the precise dates of these events, from Watson's memoir]. They had a theory of what the structure should be, and almost certainly they had already realized the implications of this structure, as they remarked in their first paper "It has not escaped our notice that the specific pairing we have postulated immediately suggests a possible copying mechanism for the genetic material." They also knew that if the structure was as they had theorized, then the diffraction pattern should display a number of key signatures—the regularly spaced layer lines, the "X" arrangement of their intensities, and the meridional reflections—that would provide both qualitative and quantitative confirmation of the theory. Thus you should be able to imagine their excitement when they saw the clean X–ray diffraction pattern from hydrated DNA, the famous photograph 51 taken by Ros-

alind Franklin, Fig 173. As far as one could tell, the proposed structure was right.

Problem 181: Discrete helices, more generally. Show that most of what was said above can be generalized to an arbitrary discrete helix, without assuming that the density is concentrated at $r = R$. That is, use only the symmetry defined by Eq (A254) for $d = nd_0$.

Problem 182: Fibers vs. crystals. We have discussed the diffraction from a helix as if there were just one molecule, and we have not been very precise about the different between amplitudes and intensities. Show that if there are many helices, all with their $\hat{z}$ axes aligned but with random positions and orientations in the $\hat{x} - \hat{y}$ plane, then the diffraction intensity from the ensemble of molecules depends only on the structure of the individual helices, and that all directions for the vector $\vec{\mathbf{q}}_\perp$ are equivalent.

It is crucial to appreciate that, contrary to what is often said in textbooks, it was not possible to "determine" the structure of DNA by looking at diffraction patterns like those in Fig 173. On the other hand, if you thought you knew the structure, you could predict the diffraction pattern—in the regime where it could be measured—and see if you got things right. This difference between experiments that support a theory, or which find something that a theory tells us must exist, and experiments that "discover" something unexpected or genuinely unknown is an incredibly important distinction, often elided.

So much has been written about this moment in scientific history that it would be irresponsible not to pause and reflect. On the other hand, I am not a historian. So let me make make just a few observations. Most importantly, I think, the story of the DNA structure combines so many themes in our understanding of science and society (separately and together) that is has an almost mythical quality, and as with the ancient myths everyone can see something that connects to their own concerns. There is the enormous issue of gender in the scientific community, something for which we hardly even had a vocabulary until decades after the event. There are the personalities of all the individuals, both as they were in 1953 and as they developed in response to the world–changing discovery in which they participated. There is the tragedy of Franklin's early death. There is the competition between Cambridge and London, and the impact of an American interloper on these very British social structures. Finally, there are issues that are more purely about the science, such as the interaction between theory and experiment, physics and biology. We could wander in this part of history for a long time. I need to come back and see what is essential, and what can be skipped. For now, let's move on.

In order to actually **determine** the structure of a large molecule by X–ray diffraction, we need to form crystals of those molecules. Crystals of a protein are not like crystals of salt or even small molecules. They are quite soft, and contain quite a lot of water. The bonds between proteins, for example, in a crystal are much weaker than the bonds that hold each protein together. On the one hand this makes growing and handling the crystals quite difficult. On the other hand, it means that the internal structures of the protein in the crystal is more likely to be typical of its structure when free in solution.

We recall that being a crystal in three dimensions means that there are vectors $\vec{\mathbf{a}}$, $\vec{\mathbf{b}}$, and $\vec{\mathbf{c}}$ such that the density is the same if we translate by integer combinations of these vectors,

$$\rho(\mathbf{x}) = \rho(\mathbf{x} + n\vec{\mathbf{a}} + m\vec{\mathbf{b}} + k\vec{\mathbf{c}}). \tag{A272}$$

This means that the density can be expanded into a Fourier series,

$$\rho(\mathbf{x}) = \sum_{knm} \tilde{\rho}_{knm} \exp\left[i(k\vec{\mathbf{G}}_a + n\vec{\mathbf{G}}_b + m\vec{\mathbf{G}}_c)\cdot\vec{\mathbf{x}}\right], \tag{A273}$$

where the $\vec{\mathbf{G}}_i$ are the "reciprocal lattice vectors." As a result, the X–ray scattering cross–section is a set of delta functions or "Bragg peaks,"

$$\sigma(\vec{\mathbf{q}}) \propto \sum_{knm} |\tilde{\rho}_{knm}|^2 \delta(\vec{\mathbf{q}} - k\vec{\mathbf{G}}_a - n\vec{\mathbf{G}}_b - m\vec{\mathbf{G}}_c). \tag{A274}$$

Problem 183: Details of diffraction. Fill in the details leading to Eq (**??**), including the relationship between the reciprocal lattice vectors $\vec{\mathbf{G}}_i$ and the real lattice vectors $\vec{\mathbf{a}}$, $\vec{\mathbf{b}}$, and $\vec{\mathbf{c}}$.

Even if we can make a perfect measurement of $\sigma(\vec{\mathbf{q}})$, we only learn about the magnitudes of the Fourier coefficients, $|\tilde{\rho}_{knm}|^2$, and this isn't sufficient to reconstruct the density $\rho(\vec{\mathbf{x}})$. This is called the phase problem. For small structures it is not such a serious problem, since the constraint that $\rho(\vec{\mathbf{x}})$ has to built out of discrete atoms allows us to determine the positions of the atoms from the diffraction pattern. But for a protein, with thousands of atoms in each unit cell of the crystal, this is hopeless.

The phase problem was solved experimentally through the idea of "isomorphous replacement." Suppose that we could attach to the each molecule in the crystal one or more very heavy atoms, in well defined (but unknown) positions. If we can do this without disrupting the packing of the molecules into the crystal, then the positions of the Bragg peaks will not change, but their intensities will. If we can approximate the density profiles of the

heavy atoms as delta functions (which should be right unless we look at very large $|\vec{\mathbf{q}}|$), then

$$|\tilde{\rho}_{knm}|^2 \to \left| \rho_{knm} + \sum_\mu Z_\mu e^{i\vec{\mathbf{q}}_{knm}\cdot\vec{\mathbf{x}}_\mu} \right|^2, \tag{A275}$$

where $\vec{\mathbf{q}}_{knm} = k\vec{\mathbf{G}}_a - n\vec{\mathbf{G}}_b - m\vec{\mathbf{G}}_c$, Z_μ is the charge of the μ^{th} heavy atom and $\vec{\mathbf{x}}_\mu$ is its position. In the simple case of one added heavy atom, we can choose coordinates so that its position is at the origin, and then it should be clear that the change in intensity on adding the heavy atom is directly sensitive to the value of $\cos\phi_{knm}$, where ϕ_{knm} is the phase of the complex number ρ_{knm}. Thus, one needs at least two different examples of adding heavy atoms to determine the phases unambiguously.

Do we need to say more here? Show in detail how two replacements determines the phase? Give a problem? I honestly don't know if one has to rely on absolute measurements, as one might think naively from the equations ... check!! Say something about other approaches to the phase problem.

The density really consists of discrete blobs corresponding to atoms, and—if we can look at sufficiently high resolution—additional density in the bonds between atoms. For the moment let's think just about the atoms. Then the density has the form

$$\rho(\vec{\mathbf{x}}) \approx \sum_\mu f_\mu \delta(\vec{\mathbf{x}} - \vec{\mathbf{x}}_\mu), \tag{A276}$$

where $\vec{\mathbf{x}}_\mu$ is the position of the μ^{th} atom and f_μ is an effective charge or scattering density associated with that atom. Thus the scattering cross–section behaves as

$$\sigma(\vec{\mathbf{q}}) \sim \sum_{\mu\nu} f_\mu f_\nu e^{i\vec{\mathbf{q}}\cdot(\vec{\mathbf{x}}_\mu - \vec{\mathbf{x}}_\nu)}. \tag{A277}$$

Importantly, the positions of atoms fluctuate. The time scale of these fluctuations typically is much shorter than the time scale of the experiment, so we will see an average,

$$\sigma(\vec{\mathbf{q}}) \sim \left\langle \sum_{\mu\nu} f_\mu f_\nu e^{i\vec{\mathbf{q}}\cdot(\vec{\mathbf{x}}_\mu - \vec{\mathbf{x}}_\nu)} \right\rangle. \tag{A278}$$

If we assume that the fluctuations in position are Gaussian around some mean, then

$$\begin{aligned} \sigma(\vec{\mathbf{q}}) &\sim \left\langle \sum_{\mu\nu} f_\mu f_\nu e^{i\vec{\mathbf{q}}\cdot(\vec{\mathbf{x}}_\mu - \vec{\mathbf{x}}_\nu)} \right\rangle \\ &\equiv \sum_{\mu\nu} f_\mu f_\nu \left\langle e^{i\vec{\mathbf{q}}\cdot\vec{\mathbf{r}}_{\mu\nu}} \right\rangle \qquad \text{(A279)} \\ &\sim \sum_{\mu\nu} f_\mu f_\nu e^{i\vec{\mathbf{q}}\cdot\vec{\mathbf{r}}_{\mu\nu}} e^{-\frac{1}{2}|\vec{\mathbf{q}}|^2\langle(\delta\vec{\mathbf{r}}_{\mu\nu})^2\rangle}, \qquad \text{(A280)} \end{aligned}$$

where $\vec{\mathbf{r}}_{\mu\nu} = \vec{\mathbf{x}}_\mu - \vec{\mathbf{x}}_\nu$, and for simplicity we assume that the fluctuations are isotropic. What we see is that the scattering intensity at $\vec{\mathbf{q}}$ is attenuated relative to what we expect from a fixed structure, by an amount $e^{-\frac{1}{2}|\vec{\mathbf{q}}|^2\langle(\delta\vec{\mathbf{r}}_{\mu\nu})^2\rangle}$. These are called the Debye–Waller factors. Thus, although X–ray diffraction is a static method, it is sensitive to dynamical fluctuations in structure, although it can't really distinguish between dynamics and static disorder in the crystal.

Need to come back and see what else needs to be said, given what we need in the main text. Is it worth talking about other methods, such as EM and NMR? The motifs of protein structure? ... not sure what we need or want.

You should read the classic trio of papers on DNA structure, which appeared one after the other in the April 25, 1953 issues of *Nature*: Watson & Crick (1953a), Wilkins et al (1953) and Franklin & Gosling (1953). The foundations of helical diffraction theory had been given just a year before by Cohcran et al (1952); a brief account is given by Holmes (1998). The astonishing realization that the structure of DNA implies a mechanism for the transmission of information from generation to generation was presented by Watson & Crick (1953b). It is especially interesting to read their account of the questions raised by their proposal, and to see how their brief list became the agenda for the emerging field of molecular biology over the next two decades. The rest is history, as the saying goes, so you should read at least one history book (Judson 1979).

Cochran et al 1952: The structure of synthetic polypeptides. I. The transform of atoms on a helix. W Cochran, FHC Crick & V Vand, *Acta Cryst* **5,** 581–586 (1952).

Franklin & Gosling 1953: Molecular configuration in sodium thymonucleate. RE Franklin & RG Gosling. *Nature* **171,** 740–741 (1953).

Holmes 1998: Fiber diffraction. KC Holmes, `http://www.mpimf-heidelberg.mpg.de/~holmes/fibre/branden.html` (1998).

Judson 1979: *The Eighth Day of Creation* HF Judson (Simon and Schuster, New York, 1979).

Watson & Crick 1953a: A structure for deoxyribose nucleic acid. JD Watson & FHC Crick, *Nature* **171,** 737–739 (1953).

Watson & Crick 1953b: Genetical implications of the structure of deoxyribonucleic acid. JD Watson & FHC Crick, *Nature* **171,** 964–967 (1953).

Wilkins et al 1953: Molecular structure of deoxypentose nucleic acids. MHF Wilkins, AR Stokes & HR Wilson, *Nature* **171,** 738–740 (1953).

Need classic refs about protein structure and crystallography; more if we do more.

---

6. Berg and Purcell, revisited

In the spirit of Berg and Purcell's original discussion, the simplest example of noise in a chemical system is just to consider the fluctuations in concentration as seen in a small volume. To treat this rigorously, let's remember