DNA Structure (reduced version)*

Outline:

A. Overview
B. DNA world prior to the double helix
C. The double helix
D. Directionality and palindromic sequences

A. Overview

The information contained within DNA determines the potentialities of a cell. The information within protein present within a cell determines the actuality of a cell. Between the two are enough details to fill several thick textbooks (see **Fig. 1** for a crude view of the whole). We've already considered the end product of this flow of information – proteins – and how the information in the form of an amino acid sequence can determine their structures and properties. Now we turn our attention to the beginning, to DNA. What is its structure, and how does that structure determine its properties, particularly the property of self replication?

B. DNA world prior to the double helix

DNA, or at least a phosphorus-rich compound found in nuclei, has been known since the 19th century, but even as late as Dorothy Wrinch's hypothesis in 1936 concerning chromosome structure,[1] it was thought to play a minor, purely structural role. This is understandable, considering the view of DNA structure prevalent at that time (**Fig. 2**). DNA was thought to be uniform in its composition and structure, a tetranucleotide composed of one copy of each of the four nucleotides – adenosine, cytosine, guanosine, and thymidine. Even though it soon became apparent that DNA was much bigger than a tetranucleotide, the thought remained that the large molecule might be merely a polymerization of a tetranucleotide monomer.[2] DNA was clearly a poor candidate to carry genetic information. That role was given to protein, recognized to be highly variable in structure.

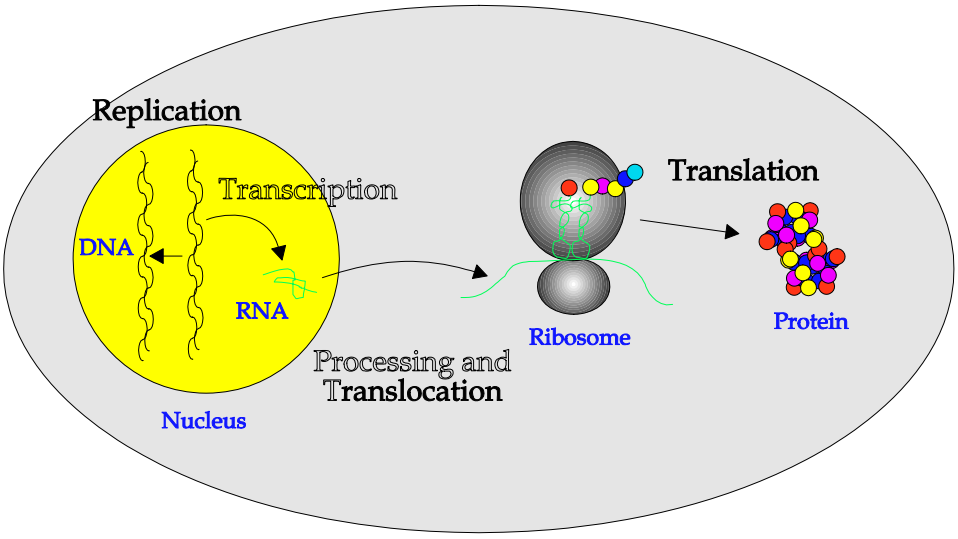

Figure 1: Flow of information from DNA to protein.

* The non-reduced version of these notes, with experimental justification for much of what is presented, can be found at http://www.people.vcu.edu/~elhaij/bnfo491/Units/DNA/DNA-structure.pdf.

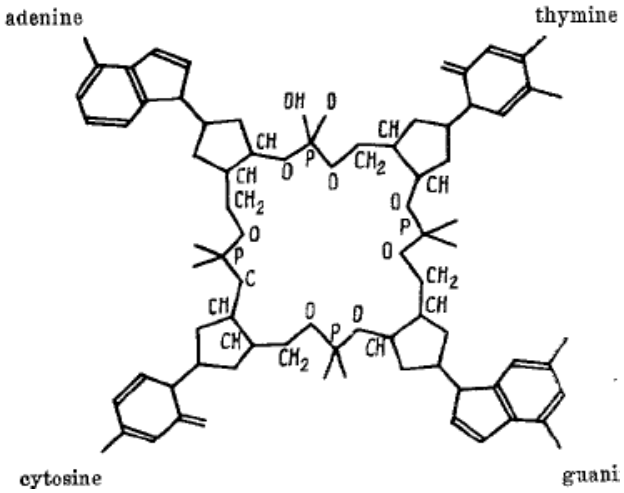

Figure 2. Proposed structures of DNA based on equal composition of nucleotides. From Reference 1.

Figure 3. Oswald Avery. http://www.chemistryexplained.com/Ar-Bo/Avery-Oswald.html

SQ1. Why is it difficult to imagine that a regular polymer consisting of repeated 4-nucleotide units might be the genetic material?

For many, work published by Oswald Avery (**Fig. 3**), Colin MacLeod, and MacLynn McCarty in 1944[3] turned the genetic world in its head. They found that a cell-free component from a pathogenic bacterium, Pneumococcus (now called Streptococcus pneumoniae), was able to transform a nonpathogenic strain to pathogenicity. The transforming component appeared to be DNA, with no discernible contamination by protein, lipid, or polysaccharide. How could a repetitive molecule, as DNA was envisioned to be, contain sufficient information to specify the production of something as complex as the pathogenic envelope capsule of the bacterium?

The article was highly influential, even though its conclusions were not widely accepted. Many worried that Avery et al's transforming principle contained contaminating protein – thought far more likely to be the true genetic material.[4] The controversy brought to mind a famous clash a decade earlier in which catalysis was claimed for preparations that were substantially protein-free, preparations that later turned out to have traces of protein, the true agent of catalysis.[5] It is also reminiscent of a more recent controversy in which many insisted that preparations of prions, the causative agents of Mad Cow Disease and other afflictions, must be contaminated with nucleic acid in order to explain their transmissibility.[6]

SQ2. What experiments can you imagine that might resolve each of the three controversies just described?

Erwin Chargaff (**Fig. 4**) was one of the few who embraced the conclusions of Avery, MacLeod, and McCarty, and he changed the course of research in his lab to test the validity of the conflicting tetranucleotide hypothesis, by analyzing the composition of DNA from many organisms.[7] To do this, Chargaff had to devise new analytical tools that were suitable for the analysis of small amounts of DNA, but once this was accomplished, experimental results came quickly. By 1951, results such as those shown in **Fig. 5** provided convincing evidence that the tetranucleotide hypothesis was not applicable to natural DNA.[8]

SQ3. What in Fig. 5 might persuade you that the tetranucleotide hypothesis was not tenable, at least in ox DNA?

Figure 4: Erwin Chargaff. National Library of Medicine.

TABLE 1

Composition of desoxyribonucleic acid of ox (in moles of nitrogenous constituent per mole of P)

CONSTITUENT	THYMUS			SPLEEN		LIVER
	Prep. 1	Prep. 2	Prep. 3	Prep. 1	Prep. 2	
Adenine	0.26	0.28	0.30	0.25	0.26	0.26
Guanine	0.21	0.24	0.22	0.20	0.21	0.20
Cytosine	0.16	0.18	0.17	0.15	0.17	0.16
Thymine	0.25	0.24	0.25	0.24	0.24	0.25
Recovery	0.88	0.94	0.94	0.84	0.88	0.87

Figure 5: Nucleotide composition from DNA of ox. This is one of many organisms studied by Chargaff's group. Numbers represent the total nucleotide identified in the preparation as a fraction of the whole. Note that the sum of the four fractions does not add up to 1.0. Each preparation represents the result from a separate experiment. Taken from Chargaff (1951), reference 10.

SQ4. Does Fig. 5 support the idea that the A:T and G:C ratios are close to one? If it isn't close enough for your tastes, what could be the reason for the discrepancy?

That was the first order of business. With the tetranucleotide hypothesis out of the way, there was no clear reason why DNA could not carry genetic information, as indicated by Avery, MacLeod, and McCarty's experiment. Chargaff's experiments led to many more interesting regularities in nucleotide composition besides the most famous ones sometimes called the Chargaff Rules, including some that whose basis is not understood even today.[9] But that's for another time.

C. The double helix

Any physical model of the genetic information must explain two remarkable features: (1) Genetic information is replicated from parent to progeny, and (2) Genetic information somehow determines the structure and function of the whole organism.

Anyone schooled in x-ray crystallography would see immediately from the cross in the picture of Rosalind Franklin and Raymond Gosling (**Fig. 6**)[10] that the compound that produced the image must be related to a helix.[†] From a measurement of the distance between the spots in the cross, one could calculate from simple trigonometry that one turn of the helix was 3.4 nm.

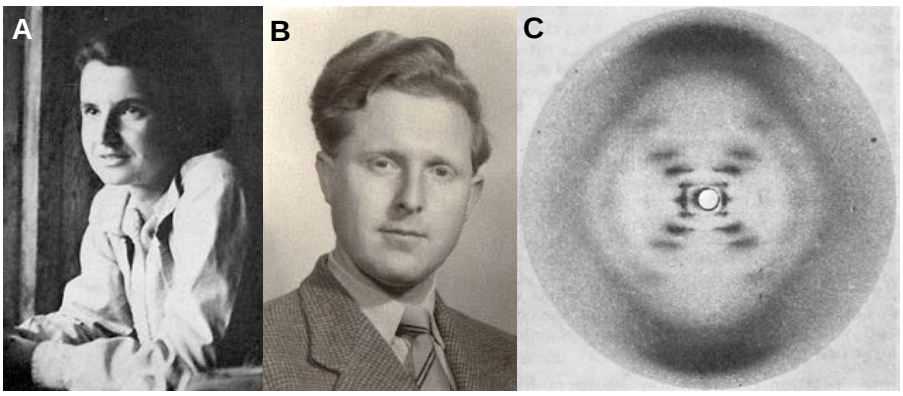

Figure 6. X-ray photograph of DNA.
(**A**) Rosalind Franklin,
(**B**) Raymond Gosling,
(**C**) DNA form B, from Franklin & Gosling (1953), reference 14.

[†] It doesn't take too much schooling. You can see how this and subsequent conclusions were drawn by looking at the non-reduced notes at http://www.people.vcu.edu/~elhaij/bnfo491/Units/DNA/DNA-structure.pdf.

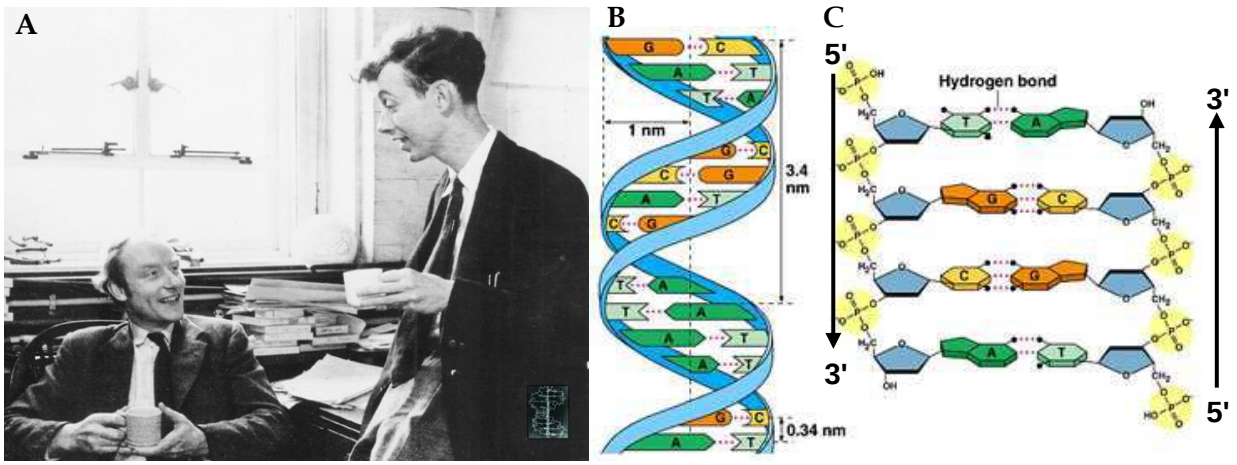

Figure 7. Watson and Crick model of DNA. (A) Francis Crick and Jim Watson **(**Photo credit: The Warden Collection, NY). **(B)** Base pairing and dimensions of double helix. (**C**) Antiparallel strands. (Figure slightly modified from one copyrighted by Pearson Education, Inc. publishing as Benjamin Cummings)

There was another immediately obvious feature from that photograph. The diffuse dark spots at the top and the bottom of the photographs (**Fig. 6C**). These more distant reflections indicated more closely spaced repetitive elements – the spacing between consecutive units (the nucleotides). Distance to these reflections permitted a calculation of the nucleotide spacing – 0.34 nm. It followed from this and the previous measurement that there are 10 nucleotides per large repeating unit, that is per turn of the helix. These features all became incorporated in the model of DNA structure put forth in 1953 by Jim Watson and Francis Crick (**Fig. 7**).[11]

What is perhaps the most obvious feature of the model shown in **Fig.** 7C, the two strands of the helix, was not so easy to deduce. Its story is told elsewhere.[‡] For now, suffice to say that Crick recognized from the shape of the crystal noted by Rosalind Franklin that the structure of DNA must have dyad symmetry, that is, it should be the same if rotated 180°.[12] This was possible only if the two helices were antiparallel to each other. We'll discuss soon the implications of this.

The second major contribution was the idea that the helices could be held together by hydrogen bonds between complementary nucleotides (**Fig. 7B** and **7C**). The base pairing of complementary nucleotides is probably the main reason why Watson and Crick's model achieved rapid acceptance. For the first time, a model had been put forward that could suggest a mechanism by which the genetic material could be replicated – the molecule already contained within it two copies of the same information. It was much less clear how DNA could specify the structure and function of a cell, but the model made clear that the answer lay in the order of nucleotides. There was nothing left that varied.

SQ5. If one strand of DNA had the sequence 5'-GGACT-3', what would be the sequence of the second strand? Draw the double helix.

SQ6. How is it that Watson and Crick's model accounts for the replication of the genetic material?

SQ7. Could you have come up with the model yourself? No need to wonder... give it a try using *Build Your Own DNA*, available on the course web site!

‡ See the non-reduced notes at http://www.people.vcu.edu/~elhaij/bnfo491/Units/DNA/DNA-structure.pdf.

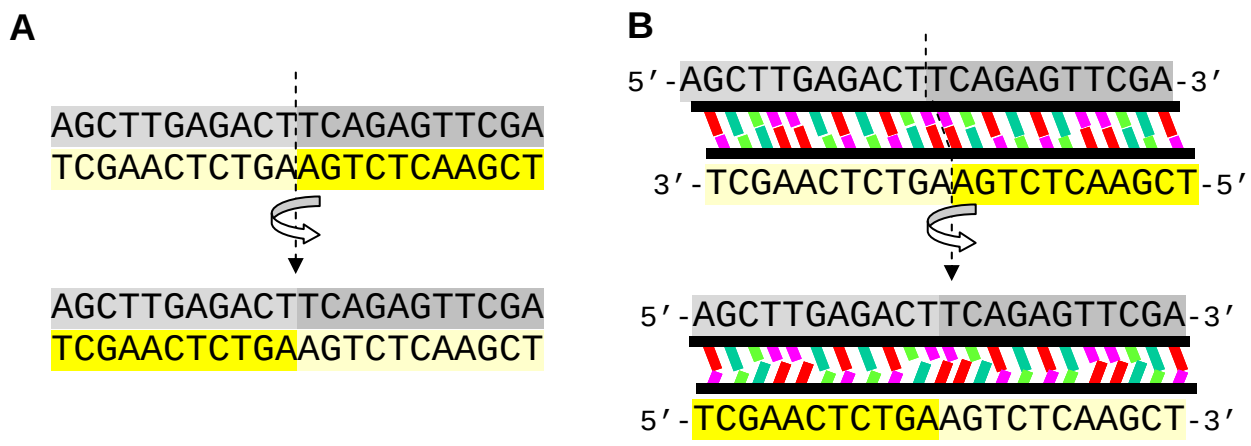

Figure 8: Directionality of DNA. Two views of a DNA sequence that appears to be palindromic (reads the same forwards and backwards). (**A**) False picture of reality, implying that the bottom strand can be rotated 180° and still base pair with the top strand. (**B**) More realistic view, illustrating that rotating the bottom strand 180° prevents the nucleotides from aligning well enough to base pair. The bottom structure thus cannot occur.

D. Directionality and palindromic sequences

The fact that DNA has directionality – 5'-to-3', top to bottom in one strand and 5'-to-3', bottom to top in the other – lays on us the responsibility that we consider this when we talk about a DNA sequence. The sequence 5'-ATG-3' is an entirely different beast from 3'-ATG-5'. To keep our sanity, those who deal with sequences have stipulated that sequences will always be considered as 5'-to-3' unless explicitly specified 3'-to-5'. So if you are given a sequence ATG, you may be sure that it what is meant is the 5'-ATG-3' version or perhaps:

5'-ATG-3'
3'-TAC-5'

The antiparallel nature of DNA has profound implications for the replication of DNA, a subject for later consideration.

SQ8. One strand of DNA has the sequence GGACT. What is the sequence of the opposite strand?

SQ9. Which of the following represent two DNA sequences that will pair with each other?

- **a. 5'-GGAGTT-3' and 5'-CCTCAA-3'**
- **b. 5'-GGAGTT-3' and 3'-CCTCAA-5'**
- **c. 5'-GGAGTT-3' and 5'-AACTCC-3'**
- **d. 5'-GGATCC-3' and 5'-GGATCC-3'**

Nucleic acid sequences that can basepair with themselves are called "palindromes" and play an important role in DNA and RNA function. Palindromes are sequences that read the same forwards as backwards. You may be familiar with palindromes from those in English, such as "Madam, I'm Adam", but don't be fooled: biological palindromes have an important difference. The sequence "AGTTGA" would be a palindrome in English but it is not a DNA palindrome, because DNA is always read 5' to 3'. Reading it backwards is not done by reading the same sequence 3' to 5' but rather by reading the ***opposite***, antiparallel strand 5' to 3'.

Consider the sequence shown in **Fig. 8A** (top strand), AGCTTGAGCTTCGAGTTCGA. It appears to read the same forwards and backwards, as does its opposite strand, TCGAACTCGAAGCTCAAGCT. Furthermore, flipping either sequence around 180° would seem to change nothing. This semblance of reality is far from the truth, as shown in **Fig. 8B**. Flipping the bottom sequence destroys the possibility of base-pairing, because a 5'-to-3' sequence cannot bind with in parallel to another 5'-to-3' sequence. 3'-TCGAA…AAGCT-5' is not the same as 5'-TCGAA…AAGCT-3'!

SQ10. With that definition, why is AGTTGA <u>not</u> a DNA palindrome?

SQ11. Make up a 6-nucleotide DNA sequence that *is* a palindrome.

Such sequences may seem to you like artificial curiosities, but in fact palindromes are important determinants of RNA structure and are also frequent targets of DNA- and RNA-binding proteins, including proteins that regulate transcription. These sequences are therefore of great importance in the bioinformatic analysis of DNA.

For example, **Fig. 9A** shows a DNA sequence that contains a gapped palindrome. The transcription of this sequence to single-stranded RNA allows the palindrome to fold up into a hairpin structure that is recognized as the termination of transcription. **Fig. 9B** shows the well-known structure of tRNA, which arises because of many self-annealing sequences, which can be thought of as gapped palindromes.

Palindromic sequences in DNA hardly ever unpair from their complementary strands and form hairpins. However it is very common for palindromic DNA sequences to function as protein binding sites, as illustrated in **Fig. 10**. Regulatory DNA-binding proteins are often dimeric (two-subunit), with each of the two subunits binding to an arm of a palindromic sequence.

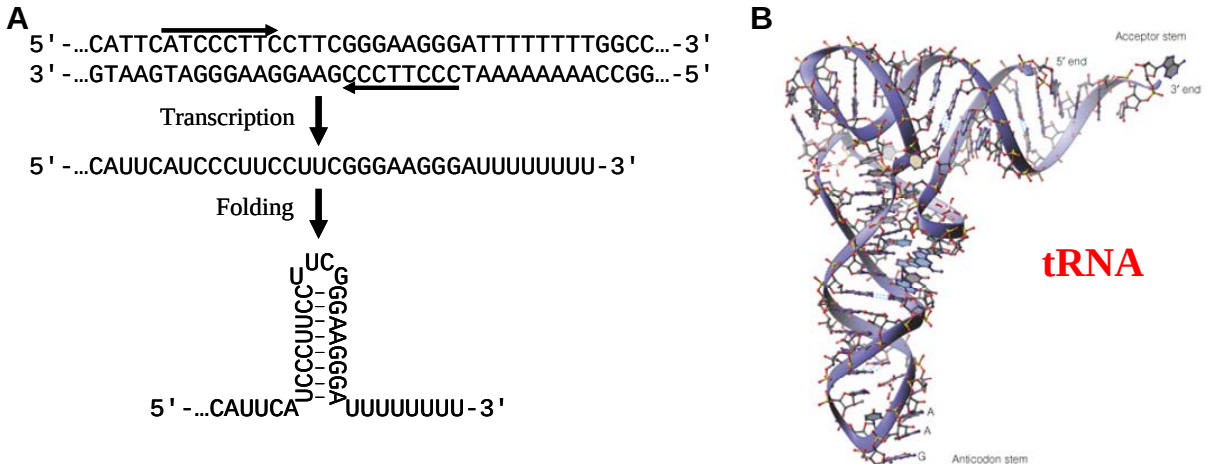

Figure 9. Palindromic RNA sequences. (**A**) Transcription of a DNA sequence containing a gapped palindrome produces an RNA that can fold to produce a common structure recognized as the termination of transcription. (**B**) RNA with many regions where self-base-pairing is possible, leading to the mature structure of a tRNA (copyright Benjamin/Cummings, an imprint of Addison Wesley Longman Inc).

There is a sound evolutionary reason why this is the case. Regulatory proteins must bind to specific sequences in order to regulate specifically one gene and not another. It takes time – lots of time – for a protein to evolve to bind to a specific DNA sequence, and it ought to take more time the longer the sequence is to which the protein must bind. That means that it would take less time for a protein to evolve that binds to a 7-nucleotide sequence than a 14-nucleotide sequence. But if the 14-nucleotide sequence is a palindrome, then a 7-nucleotide-specific protein is all you need to bind to the full 14 nucleotides. It's true, evolving a 14-nucleotide palindrome takes more time than a 7-nucleotide sequence, but less than you might think, because there are genomic processes that cause gapped palindromes to arise much more frequently than expected by chance.

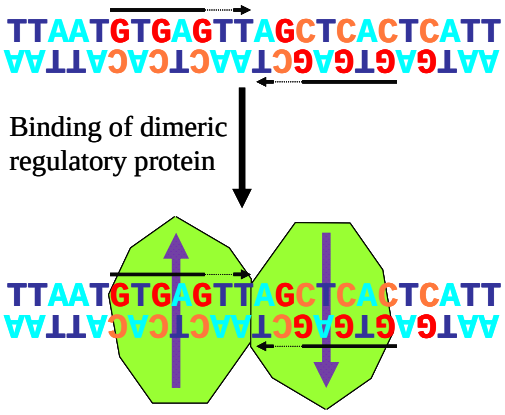

Figure 10. Binding of dimeric protein to palindromic DNA. Arrows show a region that is palindromic (dotted lines indicate a gap in the palindrome). The green blobs represent a subunit of a two-subunit regulatory protein, each subunit binding to one arm of the palindrome. The arrows indicate the opposite orientations of the subunits.

SQ12. What is the relationship between a palindromic sequence and a potential hairpin structure?

SQ13. Using arrows, identify palindromic sequences in the tRNA shown in Fig. 9B.

SQ14. Speculate on how a dimeric structure might be necessary for the functioning of a DNA-binding protein.

References

1. Wrinch DM (1936). On the molecular structure of chromosomes. Protoplasma 25:550-569.
2. Manchester KL (2007). Historical Opinion: Erwin Chargaff and his 'rules' for the base composition of DNA: why did he fail to see the possibility of complementarity? Trends Biochem Sci 33:65-70. (http://dx.doi.org/10.1016/j.tibs.2007.10.009)
3. Avery OT, MacLeod CM, McCarty M (1944). Studies on the chemical nature of the substance inducing transformation of Pneumococcal types: Induction of transformation by a desoxyribonucleic acid fraction isolated from Pneumococcus type III. J Exp Med 79:137-158. (http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2135445/)
4. Lederberg J (1994). The transformation of genetics by DNA: An Anniversary Celebration of Avery, MacLeod, and McCarty (1944). Genetics 136:423-426. (http://www.genetics.org/content/136/2/423)
5. Sumner JB (1935). Enzymes. Annu Rev Biochem 4:37-58. (http://www.annualreviews.org/doi/abs/10.1146/annurev.bi.04.070135.000345)
6. Prusiner SB (1998). Prions. Proc Natl Acad Sci USA 95:13363-13383. (http://www.pnas.org/content/95/23/13363.long)
7. Kresge N, Simoni RD, Hill RL (2005). Cohen S (2002). Chargaff's Rules: the work of Erwin Chargaff. J Biol Chem 280:172-174. (http://www.jbc.org/content/280/24/e21.full)
8. Chargaff E (1951). Some recent studies on the composition and structure of nucleic acids. J Cell Physiol Suppl 1:41-59. (http://onlinelibrary.wiley.com/doi/10.1002/jcp.1030380406/pdf)
9. Rudner R, Karkas JD, Chargaff E (1968). Separation of B. subtilis DNA into complementary strands. III. Direct analysis. Proc Natl Acad Sci USA 60:921-922.(http://www.ncbi.nlm.nih.gov/pmc/articles/PMC225140/)
10. Franklin RE, Gosling RG (1953). Molecular configuration in sodium thymonucleate. Nature 171:740-741. (http://dx.doi.org/10.1038/171740a0)
11. Watson JD, Crick FHC (1953). Molecular structure of nucleic acids. Nature 171:737-738. (http://www.nature.com/nature/journal/v171/n4356/pdf/171737a0.pdf)
12. Judson HF (1996). *The Eighth Day of Creation*. Cold Spring Harbor Laboratory Press. p.141-142.