The A to Z of DNA

Theo Durand

Introduction

The DNA double helix is arguably the most recognizable figure of molecular biology, and the basis of every living organism. Like all biological molecules, the structure of DNA is essential to its function. However, when DNA is taught, it is presented as a rigid, uniform and structurally conserved molecule. This is not entirely true, as it turns out different structures of DNA exist in an equilibrium, and although traditional DNA is the main form, others conformations occur with surprising regularity. The form most commonly associated with DNA, the structure discovered by Watson and Crick in 1953, is one of many different formations the genetic material can take. These non-canonical forms of DNA (or non-B DNAs) have been associated with a wide range of biologically important functions within the cell. Although the existence of non-canonical forms of DNA has been known for nearly as long as the B-DNA structure (the first structure of triple stranded DNA was proposed a mere 4 years after Watson & Crick's seminal paper), their mechanism of formation, biological significance and even presence in vivo are still the source of much speculation and research. Many of these non-B conformations occur transiently during the life cycle of the genome and some under strict, non-physiological conditions: as such they have long been dismissed as oddities of no physiological effect, and thus not worthy of in-depth research. It is only with the advances in genome sequencing and the systematic databasing of gene polymorphisms responsible for disease that interest for non-canonical DNA has grown. The discovery of DNA sequences with the capacity to adopt non-B conformations near the promoter regions of oncogenes and in so-called mutational hot spots has led to renewed interest in the importance of non-B DNA in genetic instability. These sequences that can adopt non-canonical conformations are primarily made up of nucleotide repeats. There are many different types of repeats, they are significant in our understanding of non-B DNA conformations and the subject is covered in greater detail further on.

In this essay I review the current research on non-canonical DNA conformations and their putative roles in gene expression, regulation, and mutation. I discuss the morphology and attributes of these non-B DNAs and how they affect the cell machinery and as a result their role in disease. I give a brief overview of DNA repeats, and non-canonical base pairings to help elucidate the mechanism behind the formation of these structures.

What is B-DNA?

B-DNA is the type of DNA everyone is talking about when they say DNA. The classic double-helix. The good thing with B-DNA is that the basics are easy to grasp, however, there are some lesser-known characteristics of B-DNA that I need to address for the reader to understand many of the talking points in this essay.

Handedness

B-DNA is right-handed, which might sound strange given DNA's apparent symmetry. What we talk about when we talk about the 'handedness' of DNA is the direction the helix spins. When we look down at model of DNA it is clearer: in B-DNA the strands go in a clockwise direction and as a result it is said to be right-handed. As one might expect this is

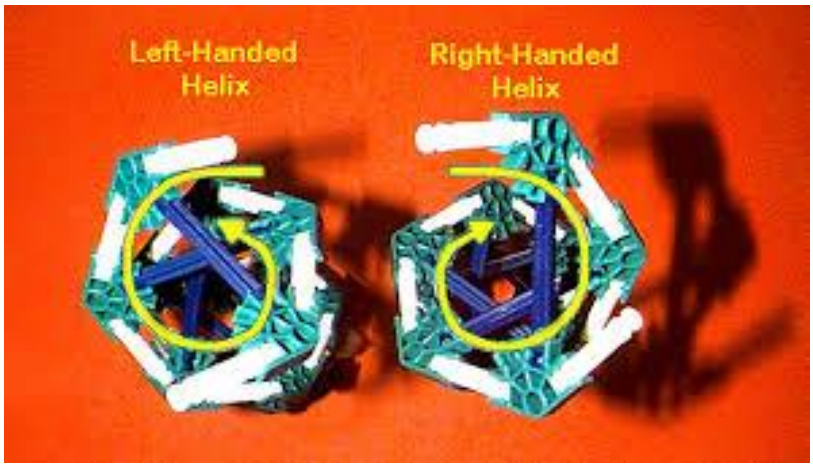


Figure 2 Model of DNA handedness

not the only conformation of DNA and we will talk about a left-handed form of DNA later. An interesting recent development is the discovery that B-DNA might be right-handed because of cosmic rays. This is significant because the handedness of DNA seemed to have no evolutionary significance (left-handed DNA is just as viable as right-handed) and has puzzled scientists since the initial structural proposal from Watson & Crick.

Grooves

Another important structural property of B-DNA is its grooves. The grooves are the spaces between the sugar-phosphate strands, this is

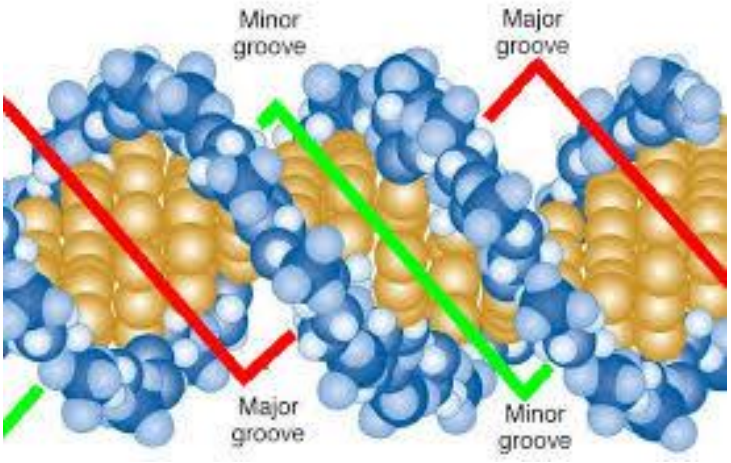


Figure 2 Model of DNA grooves

the space where DNA polymerase gains access to the DNA bases. B-DNA has 2 grooves, a major groove and a minor groove. They contain different types of information, but this is not relevant to the essay.

Non-canonical base pairing

When Watson & Crick described the structure of DNA, they also described the base pairing that would occur between the individual DNA strands. Adenine forming 2 hydrogen bonds with its compliment thymine, and guanine 3 with cytosine. However alternative base pairings have been discovered and their role in non-B DNA structures has been extensively studied.

Wobble

Wobble base pairs are a form of non-canonical base pairing that can occur in RNA during the formation of stem loops. Their importance in tRNA formation is well documented. Wobble base pairs form 2 hydrogen bonds and are as stable as Watson & Crick Base pairs.

The main Wobble bps are I-C, I-U, I-A and G-U

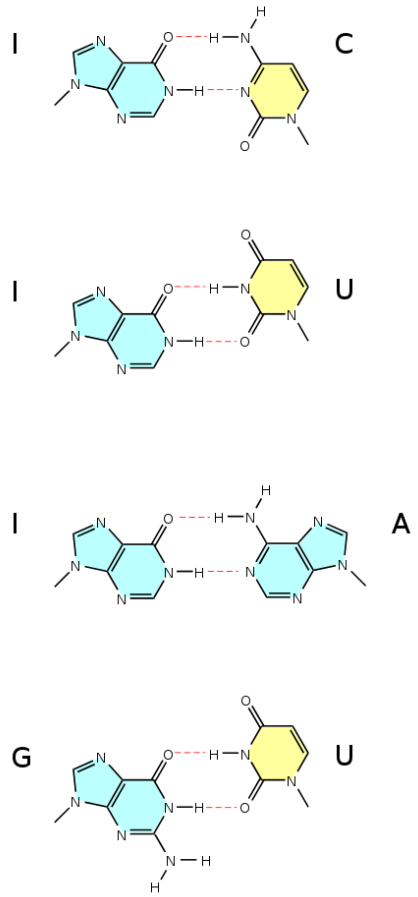


Figure 3 Wobble base pairs. The A-I pair is unusual due to it being purine-purine

where A is adenine, C, cytosine, U, uracil, and I for inosine. Inosine is the name of the nucleoside formed by the bonding of the base hypoxanthine with a ribose sugar (the shorthand I is used in keeping with naming conventions although not technically correct because inosine is a nucleoside and not a base). Inosine is formed from the deamination of an adenine nucleotide during RNA editing. The formation of uracil is the result of a homologous deamination of cytosine.

There are a large number of biological functions of Wobble base pairs, but they do not have a determined role in non-B DNA conformations due to their dependence on RNA specific nucleotides.

Hoogsteen

10 years after Watson & Crick, Karl Hoogsteen identified alternative base conformations in crystalline DNA samples. The numerous geometrical arrangements and properties of Hoogsteen base pairs can get complicated and requires knowledge of DNA isomers. To avoid getting bogged down in the details I will summarize the properties of Hoogsteen base pairs that are relevant to the understanding of non-B DNA.

- Hoogsteen base pairing can occur between complementary bases (such as A-T) but the nucleotides do not have the same spatial arrangement.
- Some Hoogsteen base pairs form preferably in acidic conditions because the base needs to be protonated. Some of these protonated arrangements form more stable bonding than traditional Watson & Crick base pairs.

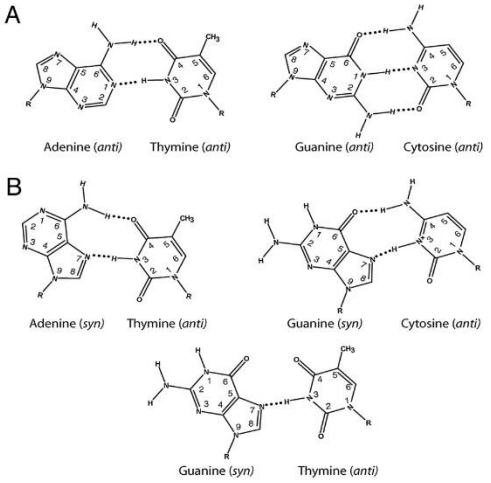


Figure 4 (A) Watson & Crick bps (B) a few Hoogsteen bps

- Hoogsteen base pairing can form triads. In these triads, 2 strands of DNA are in a Watson Crick configuration while another strand is able to bind to the bases of one strand through Hoogsteen base pairing. This is the basis for the formation of triple stranded or H DNA. (triads require a relatively stringent repeat sequence, discussed later).

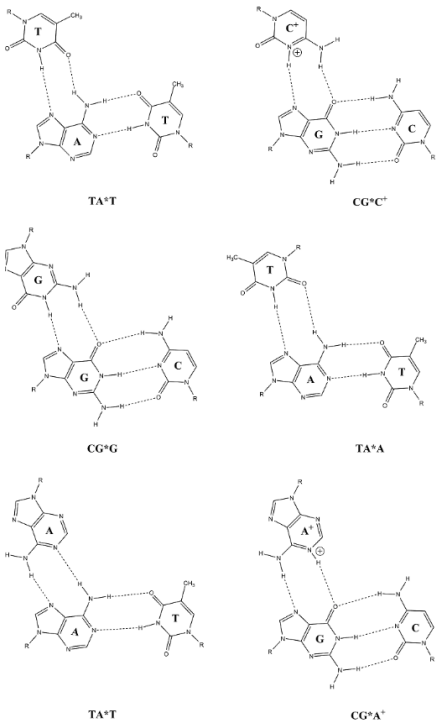


Figure 5 example of Hoogsten triads

- Hoogsteen base pairs can form a complex called a G-tetrad. In G tetrads, 4 guanine molecules bond together in a planar arrangement. Each guanine molecule forms 2 hydrogen bonds with its neighbour resulting in 8 total hydrogen bonds. These occur in 4-stranded DNA.

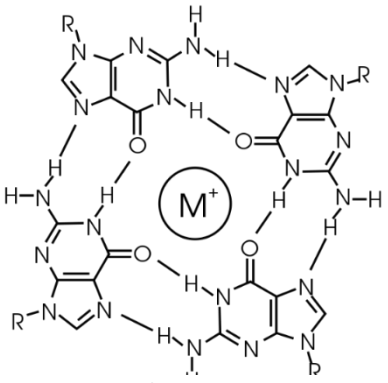
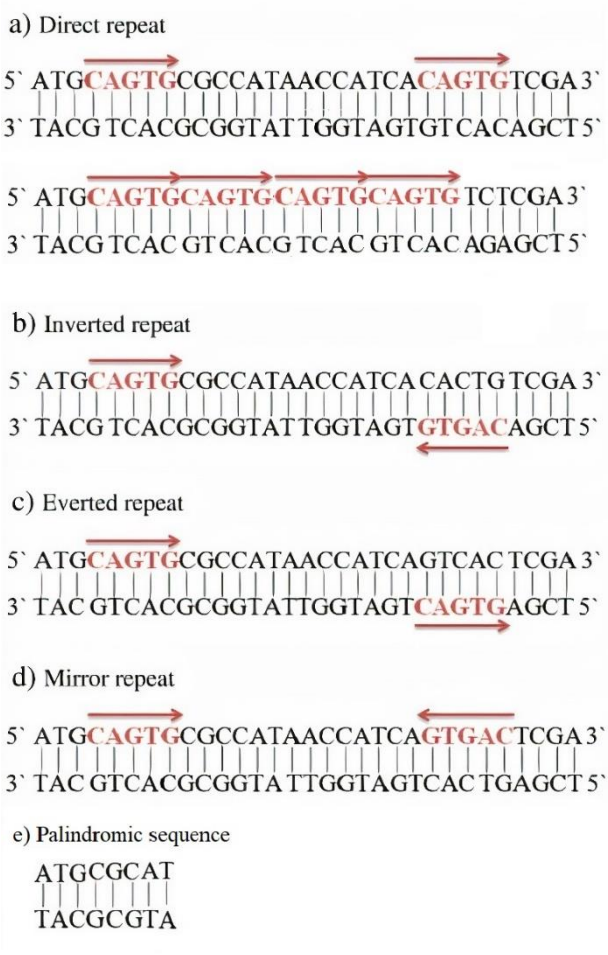


Figure 6 A G-tetrad

DNA repeats

The human genome is made up of about 50% repeat sequences. These stretches of DNA do not code for proteins: as a result their physiological function has long been dismissed. The widespread use of the term “junk DNA” to described them seems to imply that they are simply clutter in the genome. However more recent advancements in genome sequencing seem to make the picture more complex. Repeat sequences have been found near promoter regions, expansion of repeating elements can lead to disease, and, for our purposes repeats play an important role in the formation of non-canonical DNA.



Direct

Direct repeats are the most intuitive kind of repeat. They consist of the reuse of a sequence of nucleotides in one strand of genetic material.

They can be interspersed, where there is more genetic material between two repeats in between the repeats (ai) or tandemly repeating, where the repeats occur one after the other without interruption (aii).

Inverted

Where a repeat sequence in one strand is mirrored in the other strand (b).

Mirror

Where a repeat sequence in one strand is mirrored in the same strand (d).

Palindromic

A nucleotide sequence is said to be palindromic when it can be read in one direction in one strand and the converse direction in the other strand and stay the same.

E.g. In (e) we read ATGCGCAT in the top strand in the 5’ → 3’ direction and also ATGCGCAT in the 3’→ 5’ in the bottom strand.

When palindromes repeat one after the other (or interspersed like direct repeats) it is called a palindromic repeat.

A-DNA and Z-DNA

> *'The structure is an open one, and its water content is rather high.*
>
> *At lower water contents we would expect the bases to tilt so that the structure could become more compact.'*
>
> *Watson & Crick (1953)*

A and Z DNA are the two major double stranded non-B DNA conformations (C-DNA, D-DNA, E-DNA and P-DNA being some of the others). To an unfamiliar observer, they could easily be mistaken for B-DNA, however they have a few characteristics that make them different.

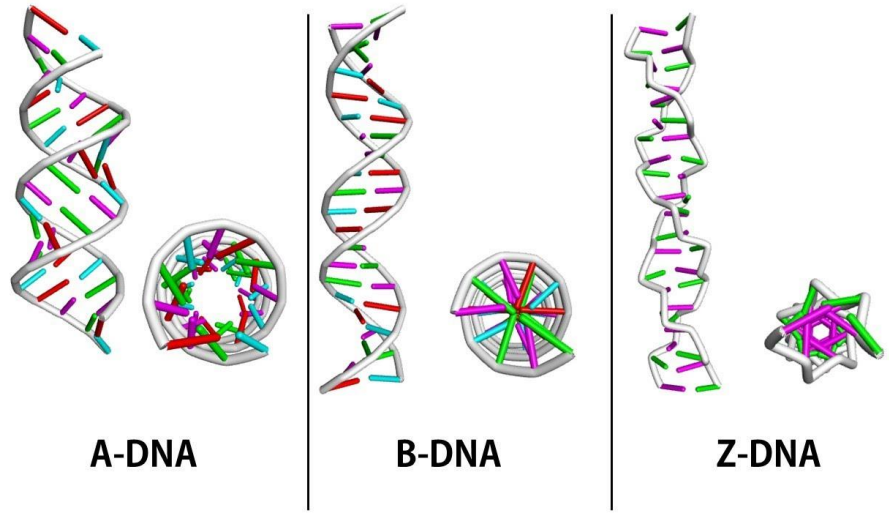


Figure 7 Comparison between the different conformations

A-DNA

A-DNA is superficially structurally similar to B-DNA. It is right-handed, does not require specific repeating sequences and has major and minor grooves. The main difference between the two is the angle the bases make with the sugar phosphate backbone. In B-DNA the angle is close to 0°, the bases appear almost perpendicular. In contrast the base pair inclination of A-DNA is about 20°, this makes the molecule look like a stouter version of traditional DNA. But no big deal, right? The cell is a judgment free zone. Turns out, not so much. The stability of the double helix it seems is less dependent on the hydrogen bonding between bases than the coin-like stacking of the bases on to of one another. Any space available for the free movement of small polar molecules like water is likely to disturb, even anneal the double helix. This is an issue in A-DNA where the helix arrangement results in a large space between the base pairs and a wider “hole” inside the helix. Apart from the instability caused by less efficient stacking this also increases the bonding distance between bases, weakening the hydrogen bonds.

As a result, A-DNA seems to only form at low humidity. Or as Watson & Crick quite correctly put it in 1953 'The structure is an open one, and its water content is rather high. At lower water contents we would expect the bases to tilt so that the structure could become more compact.'

Its crystal structure was elucidated at 75% relative

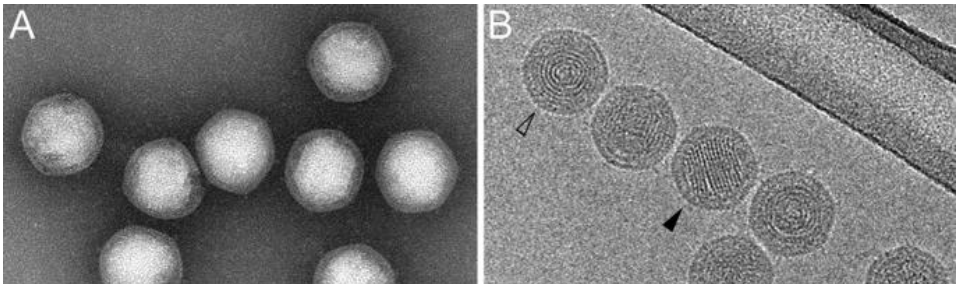


Figure 8 electron micrograph image of the SPV1 virus, which stores DNA in A form

humidity compared to 92% for B-DNA. This makes its formation highly unlikely in the necessarily aqueous environment of the cell.

Does this mean that A-DNA is a purely artificially induced form of DNA with no uses in nature? Developing research in the field of icosahedral viruses that infect hyperthermophilic archaea (a type of single celled organism that have adapted to survive in extreme conditions, in this case, boiling lakes of acid) suggest otherwise. They found that these viruses were able to store their genetic information in the A form of DNA. This is an interesting development because one of the aspects I have not yet covered in relation to A-DNA is its storage potential. The compact nature of A-DNA makes it shorter than B-DNA and means that it can pack genetic information into smaller spaces and with greater efficiency. A-DNA is also more stable at high temperatures than B-DNA, even at relatively high levels of humidity. It seems little is known about organism living in the most extreme environments on earth. As one paper put it *"A-DNA may be the prevalent storage form of DNA in everything from bacterial spores to viruses inhabiting the most extreme environments."* Although I did look into DNA storage in bacterial spores and it seems they have their own form of not-quite B-DNA (or at least the type discussed in this paper does)*. "… thus implying that DNA conformational state within the DNA-SspC complex remains similar to the canonical B form. Moreover, it has been shown that the altered photochemical properties of DNA molecules within DNA-SspC filaments cannot be straightforwardly ascribed to an A-like DNA conformation, but, at least partially, to DNA dehydration"*.

Z-DNA

Z-DNA is a left-handed form of DNA formed preferably in dinucleotide GC repeats, sometimes referred to as poly(dGC)$_2$. The symmetrical base repeats give the strand an unusual structure. In normal DNA structure the base alignment is defined by the position of the phosphate backbone; in Z-DNA however the base pairs (and other things) induce a zig-zag (hence the name) structure in the backbone. The formation of Z-DNA is not thermodynamically favourable in cellular conditions, so Z-DNA conformation can be induced by processes that release free energy within the cell. "Formation of Z-DNA removes negative supercoiling in the surrounding region of the DNA, and the energy of supercoiling stabilizes the Z-DNA conformation". Supercoiling is the action of unwinding or winding up DNA, where positive supercoiling is the action of over winding the DNA (e.g. histone deacetylation) and is an energy-necessitating process. Conversely negative supercoiling is the under-winding of DNA (histone acetylation) and releases energy.

As opposed to A-DNA, Z-DNA has been identified in vivo and hypothesized to play a number of physiologically significant functions in the cell. Repeat sequences with the capacity to adopt Z conformation have been found near gene promoter regions, especially onco- and proto onco-genes (genes with roles in cancer development), and near sites of common genetic mutation (such as chromosomal breakpoint hotspots[1]). These characteristics are common in many non-B conformations and you will see them mentioned again later. Specific to Z-DNA however are a class of proteins, Z binding proteins (ZBP), proteins that, as their name implies, bind preferentially to Z-DNA. There are quite a few contradictory pieces of literature on the role of ZBP's so I won't explore their biological importance here, I mention them because the mere existence of proteins with preferential Z binding is

[1] Chromosomal breakage, and the subsequent attempt by the cell to repair the damage through recombination or gene translocation can lead to deleterious mutations. These Z-DNA induced mutations have been associated with cancer development. (such as leukaemia and lymphomas)

evidence for the evolutionary importance of Z-DNA.

The role of Z-DNA in disease has been the subject of many papers. Here I only review one of many of these studies.

Research on Z-DNA expression in patients suffering from Alzheimer's disease has been able to determine a correlation between disease severity/ stage of progression and the levels of Z-DNA expression in the patients hippocampus (the higher the expression, the more advanced the disease). The presence of Z-DNA forming tracks near a gene associated with AD (amyloid precursor protein (APP), presenilin and APoE) was also confirmed. The hypothetical mechanism seems to be that the formation of Z-DNA absorbs energy needed for the positive supercoiling of the effector region of the genes, allowing binding and transcription of the proteins the track contains, effectively preventing epigenetic changes responsible for switching off the genes.

Again, this is a minuscule portion of the research on A and Z form DNA, with many inaccuracies and omissions for the sake of clarity.