https://doi.org/10.1038/d41586-020-02658-x

News & views

Molecular biology

DNA-binding proteins meet their mismatch

Kale Kundert & James S. Fraser

Mismatches are alterations in DNA that prevent the bases on each strand of the double helix from aligning correctly. It emerges that mismatches can bend DNA into favourable conformations for binding by proteins.

Proteins that bind to DNA are ubiquitous in biology. The ability of these proteins to bind to specific DNA sequences with high affinity is often central to their function, and it is not uncommon for a single mutation to affect the ability of a protein to bind to DNA. It is surprising, then, to find that many DNA-binding proteins can bind more tightly to sequences that have been engineered to contain a type of single-nucleotide change called a mismatch. But that is exactly what Afek *et al.*[1] report in *Nature*.

There is a key difference between a mutation and a mismatch, even though both involve changing the identity of a nucleotide. A mutation occurs on both strands of the DNA double helix. This means that base-pairing between the DNA bases on each strand is maintained. A mismatch, however, occurs on only one strand, and so normal base-pairing is abolished. In normal base-pairing, adenine (A) bases on one strand of the DNA duplex pair with thymine (T) on the complementary strand, and guanine (G) bases pair with cytosine (C) – so a change from an A–T pair to a C–G pair is a mutation, whereas a change to A–C is a mismatch. Because mismatches are not base-paired, they can distort the overall structure of the DNA more easily than mutations can (Fig. 1).

It would be reasonable to assume that distortion of DNA would impair protein binding, but in fact it can contribute to binding specificity, through a mechanism known as shape readout. In simple terms, shape readout is the ability of proteins to indirectly recognize specific DNA sequences by their characteristic 3D shapes[2,3]. This is in contrast to their ability to directly recognize specific sequences by the characteristic chemical groups present in each base pair, a mechanism known as base readout. DNA is often thought of as having the same shape, regardless of its sequence, but shape readout works because this is not strictly true. Each sequence has a preferred set of conformations (called its conformational ensemble) and can be more- or less-easily bent in different ways. Taking advantage of this, a protein that needs to bind to a specific sequence can try to bend any sequence it encounters in a way that would be most compatible with its intended target. Because bending DNA has an energetic cost, this mechanism leads to a decrease in binding affinity.

Shape readout has a role in many protein–DNA interactions[2,3], but it has been hard to study its true energetic cost, because doing so would require perturbing the shape of a DNA molecule without perturbing its sequence. Afek *et al.* realized that mismatches, being small changes to a sequence that cause large changes in shape, offer a unique way to study this phenomenon.

Working with mismatched DNA is not trivial, especially in high-throughput situations, because many standard molecular-biology techniques implicitly assume that DNA is fully base-paired. The authors therefore developed what they call a saturation mismatch-binding assay (SaMBA), which quantifies the binding of a protein to every possible single-nucleotide mismatch in a particular DNA sequence. Briefly, they manufactured a microchip arrayed with single strands of DNA encoding every possible single-nucleotide variant of a consensus sequence. Each strand was placed at a known coordinate on the chip. They then allowed a second, complementary DNA strand to flow over the array. The complementary DNA hybridized with each of the strands printed on the chip, creating duplex DNA with every possible mismatch. Finally, the authors added fluorescently labelled protein and observed its binding to the DNA using microscopy. They ran the assay using 22 different sets of DNA arrays and proteins.

SaMBA revealed not only that it is possible for mismatches to improve DNA–protein binding, but also that it is relatively common for them to do so. About 10% of all mismatches that Afek and colleagues analysed increased the affinity with which a protein bound to that sequence, including at least one such sequence for every protein. For some proteins,

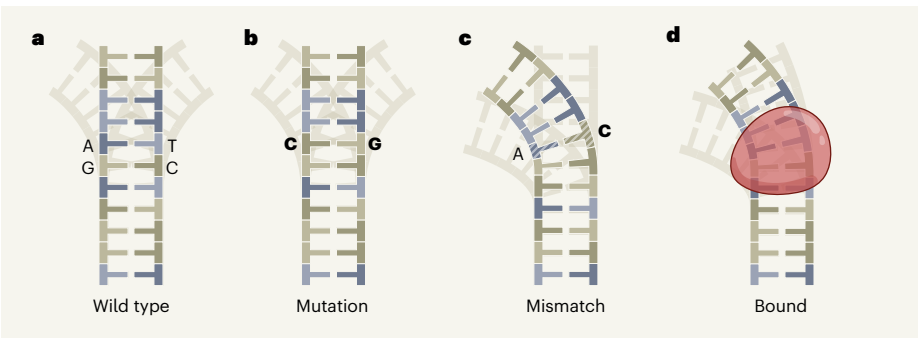

Figure 1 | Reshaping DNA. The DNA double helix involves pairs of DNA bases: adenine (A) bases on one strand pair with thymine (T) on the other, and guanine (G) bases pair with cytosine (C). **a**, A double helix can exist in a range of shapes, called a conformational ensemble. In this simple schematic, the major conformation that wild-type DNA will adopt is in the forefront, and minor conformations that it could transiently adopt are beige shadows behind. **b**, A mutation changes a pair of bases into another pair (such as A–T to C–G). This mutation is unlikely to alter the ensemble of possible conformations (although this does occasionally occur; not shown). **c**, In a mismatch, just one base is altered, disrupting base-pairing (A–T might become A–C, for instance). This disruption is likely to alter the conformational ensemble. **d**, Binding by proteins (red) can also alter DNA conformations. Afek *et al.*[1] report that 10% of mismatches bend DNA into conformations that are more similar to that of the protein-bound DNA than to the wild-type versions, making it easier for proteins to bind.

the most effective mismatch occurred in the natural target sequence, making the protein bind to that sequence even more tightly. For others, the most effective mismatch occurred in a non-target sequence, and made the protein bind to that sequence at levels comparable to those of the natural target. In both cases, the same mechanism is predominant: the mismatch pays the energetic cost of distorting the DNA so that the protein doesn't have to.

Note that to actually improve binding, the mismatch must distort the DNA in the same way as the protein would through the shape-readout mechanism. Distorting the DNA in a different way would weaken binding. The mismatch also should not interfere with any chemical contacts between the protein and the DNA — although the authors did find that mismatches can sometimes introduce favourable contacts.

Afek and colleagues' work broadens our understanding of how proteins bind to DNA, and highlights the importance of the DNA conformational ensemble in this process. In future, perhaps nucleotides that do not occur in nature[4] could be used in SaMBA to thoroughly probe the array of conformations that DNA can adopt, similarly to the way in which unnatural amino acids have been used to investigate subtle changes in protein biophysics[5]. SaMBA could also potentially be adapted to find DNA-binding proteins that are intended to bind to mismatched or chemically modified targets; such proteins would be hard to find by other means. Given that roughly one-third of transcription factors (a key class of DNA-binding protein that regulates gene expression) have no known target sequences in humans[6], this could be a productive line of enquiry.

More broadly, the finding that mismatches often improve binding might have implications for diseases such as cancer. Even a transient mismatch in the genome could prompt a transcription factor to bind in the wrong place, where it could potentially misregulate a gene and put the cell in a cancerous transcriptional state that persists even after the mismatch is repaired. Given its temporary root cause, this idea would be difficult to study or confirm. But the clear propensity for mismatches to improve binding makes such a mechanism worth contemplating.

Kale Kundert is at the Wyss Institute for Biologically Inspired Engineering, Harvard University, Boston, Massachusetts 02138, USA. **James S. Fraser** is in the Department of Bioengineering and Therapeutic Sciences, University of California San Francisco, San Francisco, California 94158, USA.
e-mail: jfraser@fraserlab.com

1. Afek, A. *et al. Nature* https://doi.org/10.1038/s41586-020-2843-2 (2020).
2. Rohs, R. *et al. Annu. Rev. Biochem.* **79**, 233–269 (2010).
3. Samee, M. A. H., Bruneau, B. G. & Pollard, K. S. *Cell Syst.* **8**, 27–42 (2019).
4. Dien, V. T. *et al. J. Am. Chem. Soc.* **140**, 16115–16123 (2018).
5. Zhang, W. H., Otting, G. & Jackson, C. J. *Curr. Opin. Struct. Biol.* **23**, 581–587 (2013).
6. Lambert, S. A. *et al. Cell* **172**, 650–665 (2018).