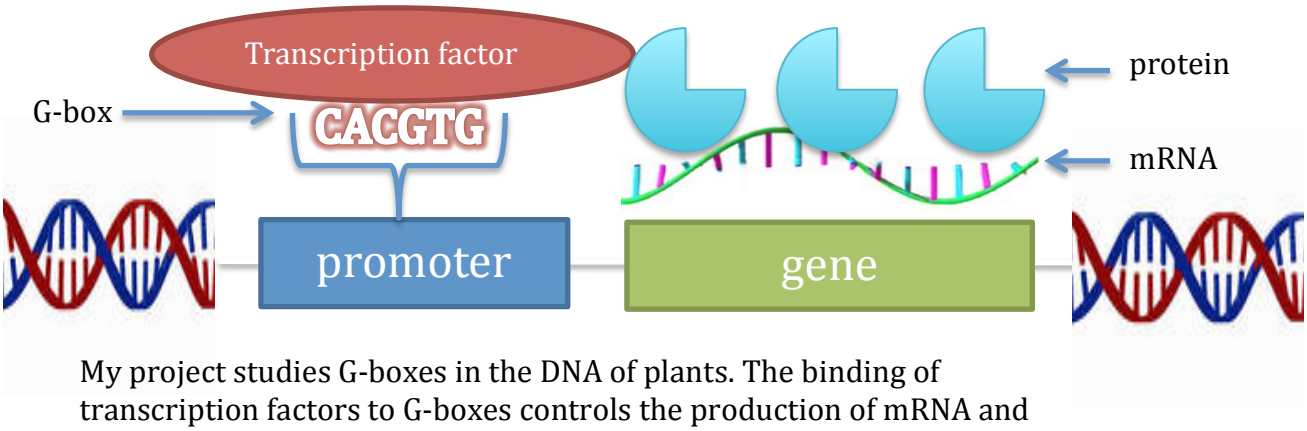
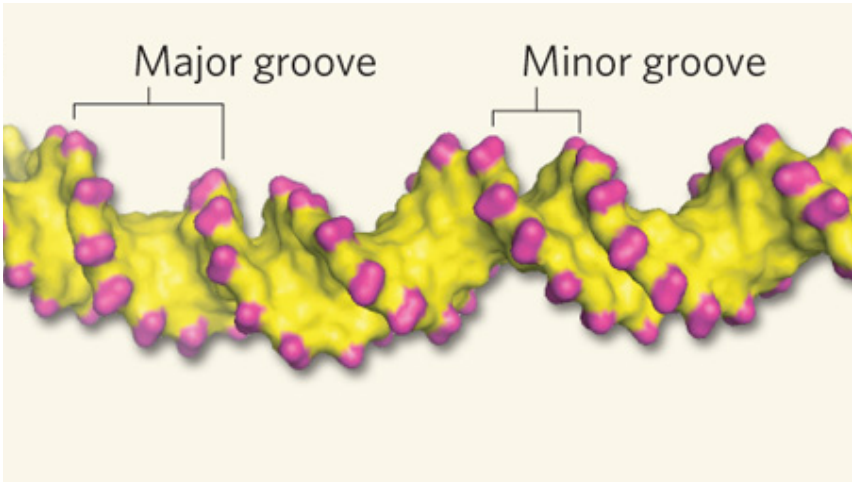
Sam Shepherd - Project at the Sainsbury Laboratory

DNA Shape of G-boxes



My project studies G-boxes in the DNA of plants. The binding of transcription factors to G-boxes controls the production of mRNA and hence of proteins, so a G-box is like the on-switch for a gene. It is little understood why certain transcription factors bind to certain G-boxes. My project investigates whether the DNA shape of G-boxes is an important factor in this.

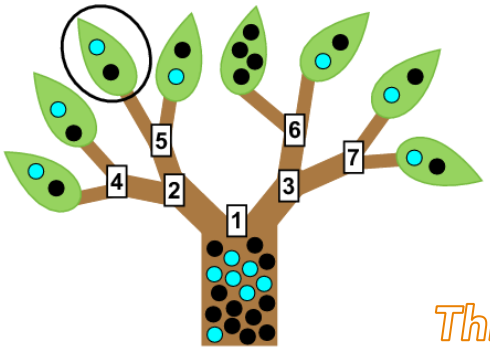


My DNA shape data included parameters like minor groove width.

I also looked at the DNA shape in the regions flanking the G-box.

I applied the machine-learning algorithm Random Forest to use DNA shape to distinguish between two groups of genes that are known to bind to different transcription factor families.

Old sequence	ATGCC	CACGTG	TTGCT
New sequence	ATGC**A**	CACG**A**G	TTGCT
Simulated new sequence	ATGC**?**	CACG**?**G	TTGCT



Where there had been base mutations in my data, I randomly simulated my own base mutations. This was to see if the actual mutations conserved DNA shape more than random mutations did – properties that are conserved in evolution usually have some important function.

This all involved a lot of coding in R!