Freedom of additional signals on genes: on the combination of DNA mechanics, genetics and translation speed

Zuiddam, M.

Summary

DNA carries various forms of information. Out of these forms of information, the most well-known is classical genetic information, which features genes that code for RNAs and proteins. This genetic code consists of sequences of bases, shorthandedly called A, T, C and G. A sequence of these bases can code for a specific protein. Throughout this dissertation we discuss what is often referred to as the second layer of information on DNA: DNA mechanics. A sequence consisting of only A's and T's will bend differently from a sequence of G's and C's. This is because the intrinsic shape of DNA, as well as its flexibility, depend on the choice of its bases. An important consequence of this mechanical layer of information is the positioning of nucleosomes.

Nucleosomes consist of 147 base pairs (bp) of DNA wrapped around a protein core, like a string around a spool. They are considered the fundamental building blocks of chromosomes and are responsible for making the DNA compact and serve to form its higher-order structures. Nucleosomes also play an important role in the regulation of DNA. They may physically restrict transcription factors (proteins that regulate transcription) from reaching transcription factor binding sites on the DNA, simply by being situated on such a site. A cell can chemically modify a nucleosome (or more precisely its 'histone tails') such that a binding site becomes accessible again. These modifications may even be inherited by the offspring of an organism, which is an example of so-called epigenetics. By either allowing or restricting access to a binding site, a nucleosome may serve as an on/off switch, of which the location is extremely important.

The location of nucleosomes on the DNA is affected by the mechanical information on the DNA. The DNA on a nucleosome must be curved in a specific way, favoured by some, disliked by other sequences. This leads to a nucleosome positioning code, often represented by the probability to find a dinucleotide step (a base followed by another base) at different locations on the nucleosome. For example, the probability to find a TT, AA, or TA dinucleotide step is highest where the minor groove of DNA faces the protein core, while finding a GC step is most likely to happen at positions where the major groove faces the protein core.

Models have been created that can reproduce but cannot explain these rules. Especially the rules concerning the GC step turned out to be counterintuitive, since its probability peaks where its deformation energy peaks, too. The first main result of this dissertation is an explanation of these rules. In Chapter 2 we approach this problem by using an analytically-tractable model that reproduces the main rules. Using the Transfer Matrix Method and a novel approximation, we demystify the rules of GC and methodetically explain the other rules.

Another approach is not to investigate these dinucleotide rules, but to look at

the malleability of long stretches of DNA to contain mechanical information. In Chapter 3 we do this by investigating the very best and very worst nucleosome-attracting sequences. We map all possible DNA sequences on a graph, weighted using the probabilistic model of Tompitak et al. [10]. By using a shortest path algorithm on this graph, we find the sequences with highest and lowest possible nucleosome wrapping energy. Using a k-shortest path algorithm we find the k-highest and lowest possible energies. The two huge advantages of this method (over other methods such as Monte Carlo simulations) are that shortest-path algorithms are in principle exact and fast.

DNA in real, living organisms does not have the luxury to only code for mechanical signals. Real DNA also needs to code for RNAs and proteins. Theoretically, these two layers of information (genetic and mechanic) can exist on the same piece of DNA. A protein consists of one or more amino acid chains. The properties of a protein depend on the type and order of amino acids in such a chain, which is encoded by so-called codons (a triplet of bases) on the DNA. In many cases, multiple synonymous codons exist that code for the same amino acid. This is called the degeneracy of the genetic code. Because of this degeneracy, multiple sequences may code for the same protein while having a diverse range of mechanical properties. As a result, a protein-coding sequence may contain mechanical signals as well as genetic information.

Using the degeneracy of the genetic code we evaluated the freedom of DNA to contain both forms of information. We do so by using graphs that contain all synonymous ways to code for the same protein. On these graphs, a shortest path algorithm provides sequences with highest and lowest possible nucleosome wrapping energy that still codes for the original protein. Furthermore we present a heuristic method to create well-positioned nucleosomes on top of a gene, and showcase this method using the genome of *Saccharomyces cerevisiae*, baker's yeast. This method, too, relies on graph representations of genes and shortest path algorithms. We investigate *all* positions on *all* protein-coding genes on yeast, and manage to create nucleosome positioning signals with single-bp resolution for 99.897% of all positions.

So far we have mentioned two layers of information on DNA. By doing so, we ignored an *additional* layer of information: translation speed. Translation speed refers to the rate at which a protein is created. This rate has important consequences for the resulting proteins, as it does not only influence the rate at which proteins can be produced, but it also affects how the amino acid chain folds during translation. This folding affects the function and fidelity of a protein. The rate at which an amino acid is added to a chain depends on the codon on the DNA (and is cell-specific and species-dependent). Even though synonymous codons code for the same amino acid, the choice of codon does have an effect on the translation speed landscape. Therefore, changing a gene while keeping the genetic code intact may lead to dysfunctional proteins. Because of this, in Chapter 4 we include the conservation of translation speed in our analysis. We present several approaches that incorporate translation speed by using the graph representations of genes introduced in Chapter 3. These approaches either use pruning, where nodes in a graph are cut out that would lead to translation speed landscapes that are too different, or alter the weights of a graph to contain translation speed as well as nucleosome energy costs.

We even use the latter approach in the context of genetically modified organisms. When one puts a protein-coding sequence of one organism in a different organism, the translation speed landscape can be very different. Altering the DNA may restore this translation speed landscape to resemble the original landscape, but this altered sequence may then have a mechanical signal very different from the original signal. Therefore, we describe a heuristic method on how to change the DNA sequence of a gene, such that, when one puts this gene in a different organism, the genetical information is conserved while the mechanical information and translation speed landscape are close to their counterparts in the original organism. This method is again powered by graphs and shortest-path algorithms.

In the final part of the dissertation we discuss how genetics and mechanics of actual organisms are multiplexed (a term that refers to having two or more layers of information on a single medium, in this case, DNA). We have shown that nucleosome signals are free to exist on top of protein-coding regions of genes, and can even coexist with translation speed signals. Genes, however, also contain noncoding parts, such as introns. Therefore, the mechanical signals may simply be encoded by noncoding parts of the genes. We introduce a classification scheme for different types of multiplexing to show which organisms encode mechanical information on top of protein-coding DNA, and which organisms use different strategies. The scheme can separate the contributions of exons (which are coding regions), introns and UTRs (which are both noncoding regions) to the nucleosome positioning signal. It also introduces the concept of intraregional signals, which occur on a single region, and interregional signals, which are signals that arise from inherent differences between regions.

We study the Transcription Start Sites (TSSs), sites known to have strong mechanical signals, of a wide range of organisms. Using our scheme we show that, for many organisms, such as fish and many plants, interregional signals dominate near the TSSs. For human and many other animals, the intron (noncoding) part of the intraregional signal dominates. For rice and other cereal grains we see that the exon (coding) part of the intraregional signal dominates. The positioning signal of rice, which is greater than that of human, shows that even coding sequences in real genomes may contain strong mechanical signals.

For rice we show that a large part of this nucleosome positioning signal can be attributed not to the choice of synonymous codons but to the choice of amino acid. We show that the amino acid sequence has a significant effect on the average mechanical landscape of rice. This makes us hypothesize that mechanical information is sometimes more important than preserving protein-coding information, and that these two types of information may even compete over the course of evolution.

We were also able to include translation speed to our analysis. Results from rice and human suggest a competition between mechanical information and translation speed signals. Summerizing, we suggest that genetics, mechanics and translation speed all three may compete with each other.