The mechanical genome : inquiries into the mechanical function of genetic information

Tompitak, M.; Tompitak M.

SUMMARY

DNA, a foundational part of all life on earth that carries genetic information, is a long, chainlike molecule: a polymer. However, it is a special type of polymer, in that its constituent monomers (the nucleotides) are not all identical. They come in four varieties, generally denoted by the letters A, T, C and G. These distinct monomers are what allows DNA to encode information in the first place: the letters play a role similar to that of the ones and zeros of the binary system that our digital devices use.

However, the distinction is not purely information-theoretical. The four possible nucleotides are necessarily distinct objects, with different physical and chemical properties. Therefore, a difference in DNA sequence is not just a difference in the encoded information, it is also a difference in flexibility, intrinsic curvature, and other elastic properties of the molecule. Such differences can have far-reaching effects.

DNA regularly needs to be deformed in nature. Most importantly, it is tightly bent in order to fit into a cell. The human genome, for instance, is about two meters long. The only reason a full copy of it fits into every single tiny cell in our body is because it is ingeniously folded up. Some parts of the DNA are energetically easier to fold than others, due to the variations in elastic properties, and as a result the sequences encoded in the DNA influence this folding. We thus find ourselves with a very rich physical system, in which the information carried by the DNA is intimately connected with its physical behavior.

DNA folding is just one example of such an interplay between the information and physics associated with DNA (although it is likely the most important one). There are two main questions to ask in the broader context. First, since DNA sequence and physical behavior are so intimately linked, does nature make use of this link? It seems reasonable to expect that evolution would also explore DNA sequences based on their physical properties, if there is some benefit to be derived from it. Secondly, how can we exploit this link ourselves? Through thoughtful manipulation of sequences, what kind of properties and behavior can we bestow upon DNA?

The five projects described in this thesis are all attempts to further our grasp on these two questions. In the first project, described in Chapter 2, we take a look at how far the intrinsic curvature of DNA can be pushed.

We design DNA sequences that lead to molecules with very strong, directionally coherent curvature, such that they form superhelical structures of their own volition. Such superhelical structures look a lot like springs, and they behave similarly, being significantly more resistant to an external force than we would expect of generic DNA.

In Chapter 3, we turn to the DNA system that receives the widest interest in the literature: the nucleosome. Folding DNA into tiny cells is assisted (in eukaryotic organisms, e.g. animals, plants and fungi) by little protein cylinders around which the DNA is wrapped. The resulting DNA-protein complex is called a nucleosome. Due to the significant bending required to wrap DNA into this complex, the affinity of a piece of DNA for the nucleosome is strongly dependent on its sequence. In Chapter 3 we see how we can go beyond a simple, scalar sequence property like nucleosome affinity.

When pulling on the two ends of a piece of DNA that is wrapped around a protein cylinder, we of course expect the DNA to be peeled away. However, it turns out that, due to the geometry of that peeling process, nucleosomes are actually kinetically protected from unwrapping due to tension: there is an energetic barrier that opposes this unwrapping, and this barrier actually becomes higher, the harder you pull. The result is that unwrapping a nucleosome requires a significant amount of force, which is good because we do not in general want the nucleosomes in our cells to fall off every time they feel some tension.

It has been shown that the way in which nucleosomes unwrap, not unexpectedly, is sequence-dependent. Specifically, we know that they can be made to unwrap asymmetrically, meaning the DNA first peels away from one side only, if the DNA sequence has better nucleosome affinity in one half than in the other.

In Chapter 3 we try to push this idea further, and we design nucleosomes with a hole in their unwrapping barrier. The result is that we can make nucleosomes that are not strongly kinetically protected, and we can make them unwrap via specific pathways of our choosing. The fact that this is possible demonstrates that the nucleosome, much like the DNA polymer, should not be considered as a single complex, but rather as a class of systems, and one nucleosome can have vastly different behavior from another.

In Chapters 4 and 5 we introduce some novel methodology. Much of the work in this thesis is built upon the Mutation Monte Carlo (MMC) method invented by Behrouz Eslami-Mossallam. The idea behind this methodology is as simple as it is powerful: take a standard physical Monte Carlo

simulation of a DNA system, and add mutations to the mix. By allowing the Monte Carlo simulation to simultaneously search the conformational space of the system and its space of possible sequences, it automatically converges on sequences that have high affinity for the system, and provides us with statistics on those sequences. In Chapters 4 and 5 we expand and enrich this methodology.

Chapter 4 first shows how the MMC method can be used to generate a bioinformatical model that approximates the sequence-dependent affinity of DNA for the system being simulated with a physical model. This new model grants us significantly enhanced reach, because its approximative nature is offset by a vast gain in computational cost. It allows us to tackle problems not remotely tractable with a detailed biophysical model like the nucleosome model of Chapter 3.

The rest of Chapter 4 is dedicated to benchmarking the new model and investigating what is needed to render it accurate. In Chapter 5 we investigate in more detail the relationship between the new method of Chapter 4 and the MMC methodology. We find that it bridges the gap between MMC simulations and SELEX experiments (a type of sequence selection experiment in which sequences are selected based on their affinity to a target, such as the nucleosome, similar in many ways to the MMC method). Along the way, we gain deeper insight into how the MMC method works, and especially into the role played by temperature in an MMC simulation.

In Chapter 6 we step into the biological realm, where we put the new model from Chapter 4 to use. Thanks to the vast gain in computational efficiency, we are able to perform genome-wide analyses of real biological DNA sequences. We focus on the promoter regions of genes of a range of organisms. These are parts of the genome, in front of the genes that contain the genetic information, that influence how often the gene is utilized in a cell. An interesting role is played here by the elastic properties of the DNA sequence, and specifically by its affinity for nucleosomes. The reason is that DNA wrapped into nucleosomes cannot be read out; the nucleosomes interfere with other machinery trying to bind to the DNA. A high or low affinity in the region where the machinery for reading out genes wants to bind therefore directly influences the expression of the gene.

Real genomes are known to encode mechanical signals around such binding sites. Yeast, for instance, a simple, unicellular organism, has DNA sequences in these regions that have poor affinity for the nucleosome, in an attempt to keep the DNA accessible for reading. The human genome, on the other hand, contains the opposite signals: sequences with high

affinity, that keep nucleosomes strongly bound at these positions. This is thought to help the genome retain some of its nucleosomes in sperm cells, in which most of the nucleosomes are removed from the DNA.

Analyzing the mechanical signals in promoter regions of around 50 organisms from across the tree of life, we find a fascinating universality: unicellular organisms are all similar to yeast, and contain signals to keep nucleosomes away, while all multicellular organisms contain signals to keep nucleosomes strongly bound. Furthermore, the strength of the signals in the latter case correlates with the complexity of the organism: mammals have sequences with very good affinity, while some simpler animals like fruit flies show more moderate signals.

Whether all these signals, and this universally observed distinction between unicellular and multicellular life, serve the same purposes they are thought to do in yeast and in humans remains to be seen.

Through these inquiries and findings, the research described in this thesis has attempted to further our understanding of both of the questions posed at the beginning of this summary. We have looked into the mechanical signals that can be found in real genomes; we have pushed the limits of the properties that DNA sequences can be made to display; and we have provided new methodology that will allow us to inquire ever further into the possibilities presented by sequence-dependent DNA mechanics.