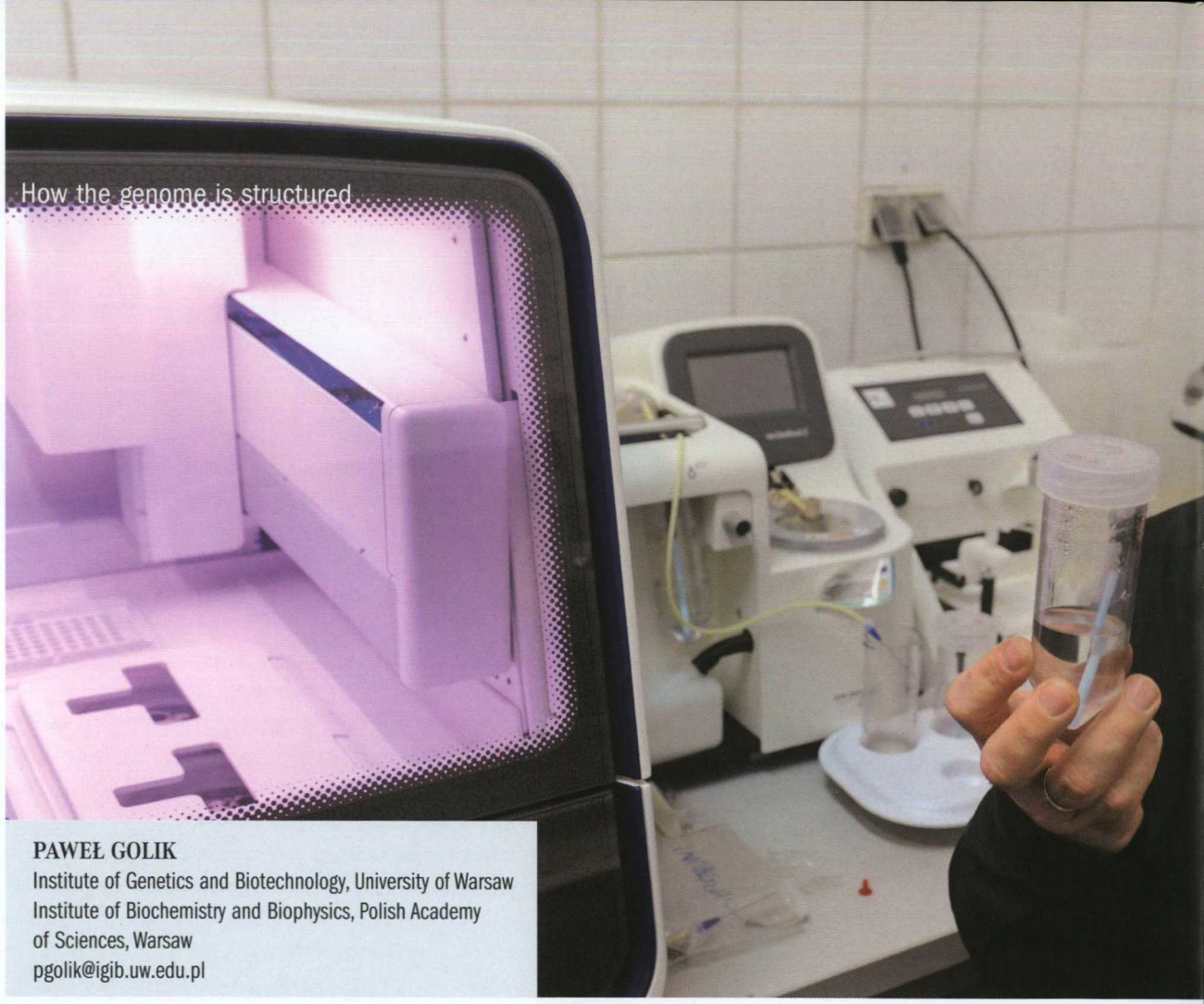
How the genome is structured



PAWEŁ GOLIK
Institute of Genetics and Biotechnology, University of Warsaw
Institute of Biochemistry and Biophysics, Polish Academy of Sciences, Warsaw
pgolik@igib.uw.edu.pl

Lego-Style Configurations

Prof. Paweł Golik talks about reading and interpreting genomes. He studies the role of nuclear encoding of proteins involved in RNA metabolism in mitochondria, and using yeast genomes in the modelling of human disorders in which nuclear control of mitochondrial genome expression is disrupted

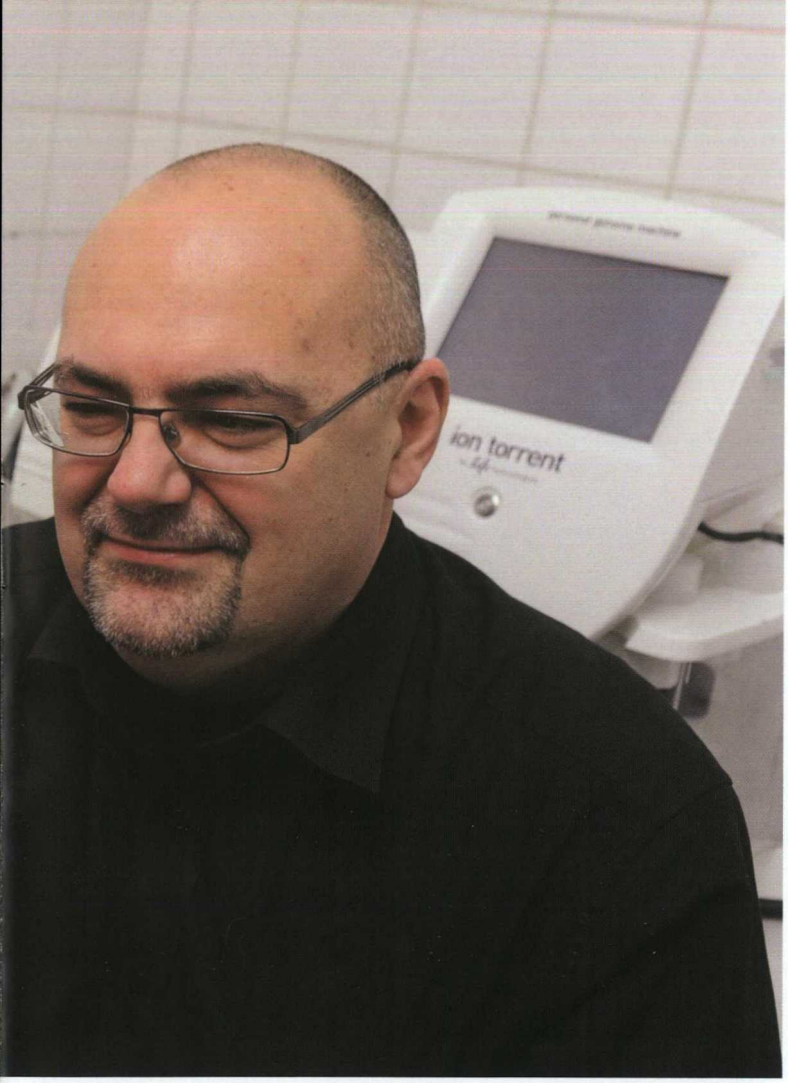
Academia: What are the fundamental levels of genomic structure?

Paweł Golik: *There are several. On the most basic level there is the primary structure – a sequence of nucleotides encoding genetic information. In simple organisms, such as bacteria and yeasts, the genome comprises genes encoding proteins or RNA, as well as featuring sequences that regulate their function. The primary structure is far more complicated in genomes of more complex organisms, such as humans; here genes comprise just a small part of DNA, while the rest of the sequence consists of vast non-coding regions. Researchers remain uncertain of their purpose. Some have termed it "junk DNA" that performs no function, while others believe that most of this non-coding DNA is involved in regulation. The truth likely lies somewhere in the middle.*

But there must also be higher-order structures.

The second level of organization is the spatial arrangement of the genome; here we are talking about structure in the more literal sense. It has long been known that DNA in cells is wound around various proteins. In complex genomes, such as human, the way it is arranged affects gene function. Cellular nuclei, which contain genomes, are highly ordered structures: specific genes are found in specific regions, and when a given gene is needed, it moves to the right part of the nucleus. The way DNA is transcribed to RNA and transported to the cytoplasm is just as ordered in specific locations.

In a traditional biochemical sense, we tend to imagine cells as vessels containing a kind of soup with individual components floating in it.



Jakub Ostałowski

That's how we used to think. Researchers used to break down cells into pieces, extract proteins, DNA or other elements, and try to recreate their functioning in test tubes. Sometimes it worked, other times not. This is probably why people imagine that components just float freely within cells. Unfortunately, treating cells this way caused information about higher order structures to be lost. It is only fairly recently that we have come to understand that cells are dense structures strictly organized on all levels – from the DNA sequence to the spatial layout in three dimensions.

To make the hierarchical organization in biology clear, I like to use Lego as an analogy. I'm talking about the old-fashioned kits which included just a few types of blocks. By arranging them in various configurations, it was possible to build almost anything your imagination desired. You didn't need to buy special blocks to create new structures or functions; you just had to arrange basic blocks in a certain way.

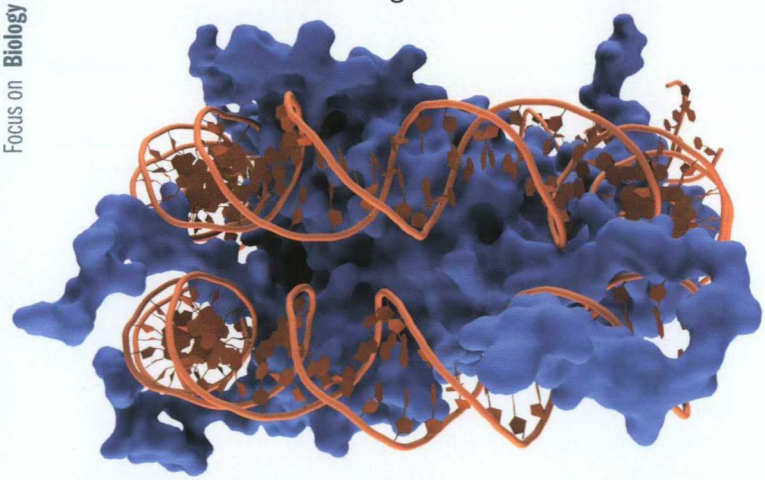
This analogy applies to many biological systems; for example, the nucleic acids DNA and RNA are built from four blocks (nucleotides) which are arranged in long sequences to create an almost limitless number of combinations. And it's the same for proteins. They are made of 20 basic amino acids; this gives so many possibilities that if we tried to synthesize all possible combinations of the amino acids found in an average protein (300 amino acids), there would not be enough atoms in the universe. The arrangement of amino acids in a protein is the primary structure, followed of course by their spatial arrangement; proteins are assembled into several typical structures such as helixes, folds, which in turn create domains. When domains are joined in different combinations, we get proteins.

How does our understanding of the structure of genomes translate into an understanding of their functioning?

We now have a reasonable understanding of the function and structure of individual genes, but there is plenty left to discover. Perhaps the greatest challenge is making a shift from describing individual genes to understanding their interactions, and the human genome comprises approx. twenty thousand genes. This is no easy task, because even such a seemingly simple trait as height depends on the interactions of 150-200 genes. And here a simple description of these genes would not suffice; we need to work with mathematicians and work out a solid theory that would help us track how they interact. We have almost boundless abilities when it comes to sequencing – or reading – genomes. But terabytes of raw data are not enough; in order to make sense of this information, we need a theoretical and systemic description, and here biology is still at the early stages.

And of course the nuclear genome is not the only source of genetic information in cells.

Eukaryotic cells are an evolutionary mosaic. Even diagrams in school textbooks make it clear they are very complex structures. They contain a number of different organelles, and some of them – mitochondria, and chloroplasts in plant cells – also contain their own genome. This has led to the hypothesis (known as the endosymbiotic theory) that organelles were once separate organisms, which long ago entered into a symbiosis with host cells. The mitochondrial genome is incomplete, since during billions of years of evolution the majority of the information it encodes has been lost or replaced by functions encoded by nuclear DNA. However, it continues to encode a handful of proteins, most of which are involved in cellular respiration. Interestingly, it is not their most important function, since many organisms respire anaerobically and are able to survive without mitochondrial DNA (e.g. petite yeast mutations). The key function of mitochondria, essential for the



Thomas Splettstoesser/Wikimedia Commons

DNA (orange) bound to histones (blue), which form a "scaffolding" for storing DNA inside the cellular nucleus

survival of all cells, is the production of iron-sulfur clusters, which are a key element of some essential cellular enzymes.

Mitochondrial genes are regulated by the nuclear genome. In humans, the mitochondrial genome comprises just 16,000 base pairs – the same number as a single average nuclear gene. This means that certain structures regulating the switching of genes on or off have been simplified, and the burden of regulation of their function falls on proteins encoded by the nuclear genome and imported from the cytoplasm. This is what my research team is studying. The mitochondrial genome evolves very fast, and the nuclear genome must keep up with these changes. In certain groups of organisms this process may be the driver behind evolution and the formation of new species.

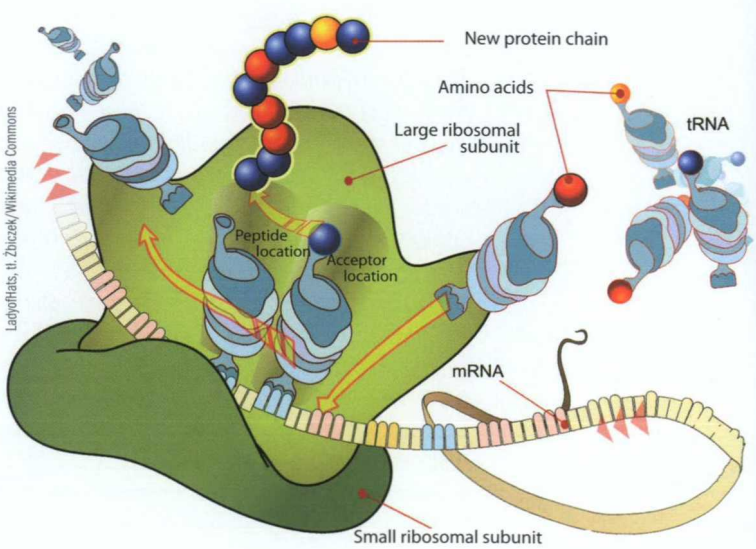
In recent years, molecular biologists have been focusing on small RNAs. What are these molecules, and what function do they perform?

Classical molecular biology stated that DNA codes for RNA, which in turn codes for proteins; Francis Crick described this as the central dogma of molecular biology. RNA does indeed perform this function, but we have since uncovered many more. In the early 1980s, researchers discovered ribozymes, and learned that DNA also performs an enzymatic function. Around the same time, the concept arose stating that when life first evolved, RNA acted as genetic material carrying information and the enzymes which executed it. It was only later that those functions became split among DNA and proteins.

The human genome contains around 20,000 genes coding for proteins, and at least 100,000 different proteins, therefore it is impossible to predict the sequence of proteins on the basis of the DNA sequence. However, using powerful computers we can do this for simple organisms. Genes comprise segments known as exons and introns. During the transcription process, introns are removed when DNA is transcribed into RNA, while exons are arranged together to form proteins. This arrangement process does not always follow the same pattern. Different fragments of the transcript can be excised as introns or remain as exons; this means that a single section of DNA may encode many different RNAs and, as a result, many different proteins.

Is this the exception or the rule?

Until recently we believed it was the exception, but we now know that over 90% of genes undergo it. This means that the definition of a gene as a fragment of DNA controlling the formation of a single product is no longer valid. A single gene can encode up to ten thousand different proteins. In order to elucidate the complexity of genetic information in animals or higher plants, we need to study not genes alone, but different RNAs and different transcripts. And of course there is still the question of gene regulation. It was once believed that this is only done by regulatory proteins, but in the late 20th century scientists discovered an entire regulatory system based on microRNA. These molecules are encoded in the regions of the genome previously regarded as junk DNA. They usually act by silencing genes or by degrading mRNA and blocking the translation process. And there are also small RNAs that modify the structure of chromatin, associated with DNA and structural proteins. Each microRNA acts on hundreds or even thousands of different genes, and each gene is controlled by many different microRNAs.



LadyofHats, tł. Zbiczek/Wikimedia Commons

Ribosomal RNA (rRNA) combines with proteins to form ribosomes, which play an enzymatic function and are used to synthesize proteins in cells through the translation process. In ribosomes, proteins are built on the foundation of an mRNA matrix from amino acids supplied by transport RNA (tRNA)

Zephyris/Wikimedia Commons

Nucleosomes are structural units of chromatin – DNA bound with proteins – which stores genetic material in the cellular nucleus

Is that the whole story of small RNA?

No. Methods of detecting trace amounts of unstable RNA reveal that the vast majority of the genome is active, and involves transcription to RNA. In the classical theory, only genes would be transcribed to RNA and the regions between them would not. Meanwhile, scientists have discovered long non-coding RNAs formed as a result of transcription of intergenic spaces. Do they perform a specific function? I believe that at least some of them may be involved in regulation, but we have much to learn.

Does the size of a genome reflect its complexity?

It was once believed that the complexity of an organism depends on the number of its genes. The human genome comprises three billion nucleotides, while bacterial genomes have just four million, so it should follow that humans are more complex because they have more DNA. And yet salamanders have a hundred times more DNA than humans, and the largest genomes actually belong to certain amoebae. As such, it seems that it is not the amount of DNA that's important, but the complexity of its control mechanisms. Some protozoa have more genes than humans since they have simpler regulatory mechanisms, so while we need just a single gene, they may need two specialized ones. In animals the systems are more refined, but that renders them more fragile, as shown by the existence of certain disorders. For example, in Down syndrome no genes are damaged. People with the condition simply have one extra chromosome, and that's enough to put the regulatory mechanism out of balance.

In plants, it's possible to vary the number of chromosomes almost freely. The majority of higher organisms are diploid, which means they have two copies of each chromosome; however, wheat has four or six, depending on variety, while calamus can have up to eight copies. This was noted by the Polish pioneer of genetics, Prof. Wacław Gajewski. When he was studying the Ranunculus *genus in the 1930s, he discovered that plants with a similar appearance may be very different in terms of the number of chromosomes.*

What are the next challenges facing biology, when we take all this into account?

Evolution takes different paths, and much depends on higher regulatory systems. The task of 21st-century biology is to elucidate them – to make a shift from describing individual building blocks to entire structures. This task is too great for us alone; we need help from mathematicians and physicists who have been studying complexity theory for far longer than biologists. And so anyone considering a career in biology must not neglect mathematics and physics, because the days of describing individual elements are coming to an end. We need to make focus on systemic descriptions, and we will not be able to do this without a solid theoretical grounding. ■

Interview by **Agnieszka Kloch**