3. DNA function: gene expression

Hello. In this lecture, we will focus on the processes that occur during gene expression. The lecture is part of module 1, Animal Genetics. The creation of this presentation was supported by the ERASMUS + KA2 grant within the project ISAGREED, Innovation of content and structure of study programs in the field of management of animal genetic and food resources using digitalization.

Genes contain biological information, but they are not capable of implementing this information in a cell themselves. Its utilization requires coordinated activity of enzymes and other types of proteins that are involved in a series of events forming gene expression.

Gene expression is usually viewed as a two-phase process that involves transcription of genetic information from a DNA sequence into its copy in an RNA sequence. The second process is translation, when the genetic information is translated from the RNA sequence into the sequence of amino acids, that is, into a protein. All of these processes are based on the structural properties of nucleic acids - namely, base complementarity and antiparallelism in the double-stranded structures of DNA.

The function of DNA is collectively described by the so-called Central dogma of molecular biology. It describes the flow of genetic information between biological molecules (i.e., nucleic acids and proteins) and is based on the knowledge of the structure of DNA as described by its discoverers Watson and Crick. Common transfers include DNA replication, RNA synthesis based on DNA (transcription), and the transfer of genetic information from RNA to protein (translation). Only specific groups of RNA viruses have the ability to replicate RNA and create DNA based on RNA, named as reverse transcription. However, it is not possible to transfer genetic information from a protein back to a nucleic acid under any circumstances.

Gene expression is a process in which information from a gene is used to synthesize a functional gene product, i.e., a protein or non-coding RNA with a specific function, and ultimately influences the phenotype. The expression of genetic information is a process that includes two steps: transcription and translation. These processes produce proteins that are the basis of correct cell and organism composition and function.

Later it was found that the process of gene expression does not only involve the formation of proteins. Only a small portion of cellular RNA, usually no more than 4% of total RNA, is composed of messenger RNA. Most of the RNA produced in a cell is used to support the process of translation or regulate gene expression, meaning they are RNA molecules that will not be translated into proteins. They thus play a direct role in the cell as RNA itself - small nuclear RNA (snRNA), small nucleolar RNA (snoRNA), small cytoplasmic RNA (scRNA), microRNA (miRNA), and possibly small interfering RNA (siRNA).

The expression scheme shows in detail most of the processes related to the expression of the structural gene, i.e., the gene that codes for a protein. In addition to the two basic processes of transcription and translation, it also includes post-transcriptional RNA modifications and post-translational protein modifications. These are a series of regulatory steps that can influence the final form of the protein coded in the DNA sequence referred to as a gene. There is also an

influence of the environment in which the cell or organism is located, which can also modify the outcome of a given expression step.

Perhaps the most important phase of transcription is the first phase, called initiation of transcription. Not all genes are constantly transcribed. On the contrary, there are mechanisms that ensure gene activation as needed. It must be said that most genes in a cell are silenced at any given moment. However, how does RNA polymerase know where the gene is in DNA located and that it should transcribe this gene? The main mechanism for initiating gene expression is the function of the gene promoter, which is a DNA sequence that initiates the transcription of a specific gene by binding to RNA polymerase and other components of the transcription apparatus, called transcription factors. RNA polymerases are enzymes that perform transcription itself - RNA polymerization based on a template strand in DNA. In eukaryotes, there are three basic types of RNA polymerases (I - III), each with typical promoters. The core promoter of RNA polymerase II in eukaryotes consists of two main segments. The first is in the region -25, TATA box, which has a conserved sequence 5'-TATAWAAR-3'. The second part of the promoter core is the initiator sequence (Inr), which, for example, is located around nucleotide +1 in mammals.

All genes undergo the first phase of gene expression, called transcription, which results in the synthesis of an RNA molecule. Transcription is a simple copying reaction. RNA is a polynucleotide, chemically differing from DNA only in that RNA has the sugar ribose instead of 2'-deoxyribose, and that thymine is replaced by the base uracil (U), which, like thymine, pairs with adenine. During the transcription of a gene, one strand of the double helix DNA serves as a template for the synthesis of an RNA molecule, and the sequence of nucleotides is based on complementarity, i.e. standard base pairing rules.

In the diagram presented on slaid 9, we can see that DNA is a double helix and genetic information is only encoded in one strand, which is referred to as the coding or positive strand. Here, the information about the amino acid sequence in the encoded protein is stored. However, due to base complementarity, the second strand, the so-called negative strand, often referred to as the anti-coding strand, serves as the template. The enzyme RNA polymerase moves along this strand and, based on base complementarity, polymerizes the growing RNA chain. The resulting primary RNA transcript carries the same information, i.e. the same sequence, as the coding positive strand, with the same orientation. The ends that are invariant within the DNA strand are designated as 5' and 3'.

The primary transcript must undergo further processing. RNA processing involves capping, splicing, and polyadenylation. Capping and polyadenylation serve to protect the RNA from degradation in the cell. Splicing specifically removes non-coding internal parts of the gene, known as introns. As an example of alternative splicing, the splicing of pre-mRNA transcript of the alpha-tropomyosin gene in rats in different types of cells is shown in the diagram. One gene can produce 9 different final transcripts, with nine slightly different but similar proteins. Light green frames represent introns; other colors represent exons. Polyadenylation signals are marked with the letter A. Dashed lines indicate the regions that have been removed by splicing.

For some genes, the RNA transcript itself is the final product of gene expression. In structural genes, the transcript is a short-term message that controls the second phase of gene expression, called translation. It involves translating the genetic information (the nucleotide sequence encoding the amino acids of a peptide) into the primary structure of a polypeptide (i.e. the

sequence of amino acids). Protein is another type of polymeric molecule completely different from DNA and RNA. In proteins, the monomers are called amino acids, and there are typically 20 different ones, each with its own specific chemical properties. Translation is based on base complementarity between the codon on mRNA (a copy of the gene) and the anticodon on tRNA (the carrier of amino acids). It utilizes the genetic code - a system of rules for encoding amino acids. While transcription occurs in the nucleus, translation occurs on ribosomes in the cytoplasm.

Pairing of codons and anticodons is a fundamental mechanism for translating genetic information into protein. The mRNA molecule brings the genetic information to the site of translation. tRNA molecules bring amino acids attached to them. Each tRNA is specific for a particular amino acid and distinguishes itself from other tRNAs by its anticodon. For example, a tRNA with the anticodon CGU always carries the amino acid alanine. And during translation, this tRNA can only bind to the GCA codon on the mRNA.

In translation, the amino acid sequence is determined by the nucleotide sequence in the mRNA. Each triplet of adjacent ribonucleotides specifies one amino acid of the protein, with the identity of the amino acid corresponding to each triplet determined by the genetic code. The genetic code is a system of rules for translation; it is essentially a translation dictionary that is universal, i.e., with a few exceptions, the same for all organisms in the world. Formally, the genetic code is usually displayed in the form of a table or graph to allow easy reading. There are 64 codons in total, of which three are nonsense (stop codons) and one is an initiation codon (also encodes the amino acid methionine). The function of a stop codon is to stop the synthesis of a polypeptide, so it must always be at the end of the coding sequence of a gene. Since there are 20 basic amino acids, the genetic code is degenerate, i.e. one amino acid can be encoded by multiple codons, but the opposite is not true, i.e. one codon always encodes the same amino acid. Note that, for example, the amino acid leucine can be encoded by 6 different codons, whereas tryptophan can only be encoded by one.

The translation process can be divided into three phases: initiation, elongation and termination. In addition to the ribosome, mRNA and tRNA, additional proteins are required for the successful completion of each phase. mRNA binds to a small subunit of the ribosome along with initiation factors. The initiation complex tRNAmet binds to the mRNA codon sequence AUG with its anticodon UAC in the peptidyl site. The large subunit of the ribosome binds to this complex. A second activated tRNA with the bound amino acid enters the A site and a peptidyl bond is formed between the two amino acids. The first tRNA without the amino acid is dissociated to form a dipeptide bound to the second tRNA.

The mRNA strand is shifted by three nucleotides, i.e. the second tRNA is shifted to the P site and the third tRNA can bind to the mRNA at the A site. In this way, the amino acid sequence is elongated exactly according to the genetic information in the mRNA. Once the stop codon in the mRNA is reached, the process is terminated, the whole complex breaks down and the protein is released. This is then further modified in various cellular organelles to produce the final 3D spatial structures that enable the final function of the protein.

And this is all for this short presentation explaining the basics of genetic information expression. Thank you for your attention.