Development of Complex Curricula for Molecular Bionics and Infobionics Programs within a consortial* framework**

Consortium leader

PETER PAZMANY CATHOLIC UNIVERSITY

Consortium members

SEMMELWEIS UNIVERSITY, DIALOG CAMPUS PUBLISHER

The Project has been realised with the support of the European Union and has been co-financed by the European Social Fund ***

**Molekuláris bionika és Infobionika Szakok tananyagának komplex fejlesztése konzorciumi keretben

***A projekt az Európai Unió támogatásával, az Európai Szociális Alap társfinanszírozásával valósul meg.

INTRODUCTION TO BIOINFORMATICS

(BEVEZETÉS A BIOINFORMATIKÁBA)

CHAPTER 1

Basic molecular biology for informaticians

(Molekuláris biológiai alpok informatikusoknak)

Péter Gál

Bioinformatics is the application of information technology in life sciences.

Generally, bioinformatics means the computer-based analysis of large biological data sets.

The area of bioinformatics is continuously expanding, since new methods for generating new types of biological data sets are emerging and improving.

The advance of high throughput data acquisition methods has fundamentally changed the biological sciences in the recent decades.

Bioinformatics is a basic, as well as an applied science.

The purpose of acquiring, storing, organizing, archiving, analyzing biological data is to draw new conclusions related to the biological systems (e.g. bacteria, plants, animal, human) and apply them in the research, biotechnology and medicine.

Thus, bioinformatics contributed to the rapid development of biotechnological and pharmaceutical industry and it has a great impact in the modern diagnostic and therapeutic methods.

Major areas of bioinformatics:

1.) Analysis of DNA sequences

2.) Analysis of RNA sequences

2.) Analysis of protein sequences

3.) Analysis of protein structures

4.) Analysis of other databases (e.g. metabolic databases, gene expression, protein-protein interactions, etc.)

5.) Other applications (e.g. drug development, protein design, personalized medicine, etc.)

Other areas are continuously emerging.

1.) DNA and RNA sequences

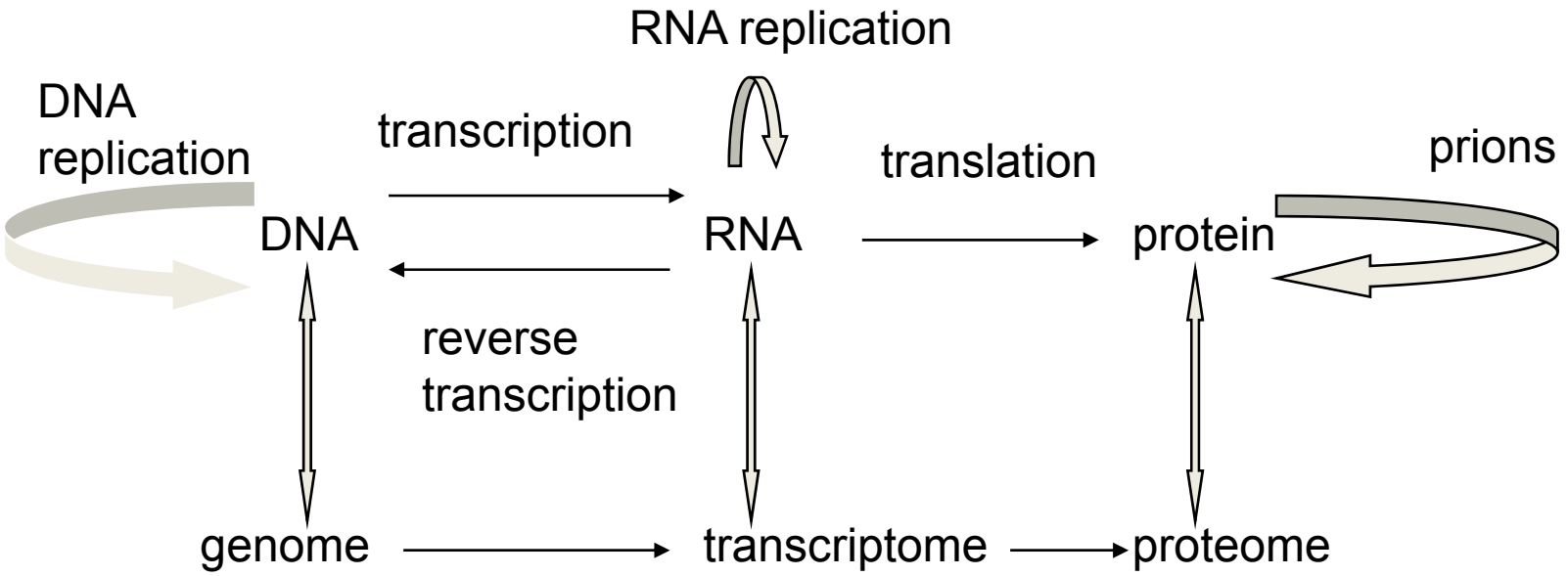
DNA, and in case of some viruses, RNA are the materials in which all the information necessary for life is stored.

DNA, RNA and protein are macromolecules (biopolymers) that carry information.

Other biologically relevant (macro)molecules such as carbohydrates and lipids cannot store information since they contain monotone repeats of one or two building units.

The primary information is the sequence of the monomeric building blocks: i.e. nucleotides in the case of DNA and RNA and amino acids in the case of proteins.

The central dogma of molecular biology tells us about the direction of information-flow between the biological macromolecules:



Genome: The genome of an organism contains all the genetic information encoded in the DNA (or RNA). In Human the genom includes the chromosomes (22 pairs of autosomes and the X and Y sex chromosomes) plus the DNA content of the mitochondrium.

Transcriptome: Transcriptome includes all the RNA molecules of a given cell or organism expressed at any given time.

Proteome: Proteome includes all the protein molecules of a given cell or organism expressed at any given time.

The information, which is encoded in a genome, transcriptome or proteome is enormous. Bioinformatics is an indispensable tool to study the recently emerged new disciplines, „omics”: genomics, transcriptomics , proteomics.

Systems biology

In the last century geneticists, biochemists and molecular biologists analyzed the properties of isolated genes and/or gene products of an organism in order to decipher the molecular basics of life. The advance of the high throughput data acquisition techniques and the bioinformatical methods made possible to analyze the function of many (preferably all) genes and/or gene produts at the same time. Now we can put the pieces of information together to form a biological system. Therefore integration is a key word in systems biology. Integration means, at the first place, integration of protein-protein and protein-nucleic acid interaction patterns within a cell or organism.

The gene

In the early 20th century the gene was defined as a part of the genetic material which is responsible for the expression of a particular feature (visible property) of an organism (phenotype). At that time the chemical nature of the genetic material was not known.

In 1940 the one gene-one enzyme hypothesis was put forward. Later it was broadened to one gene-one protein concept.

In 1944 Avery showed that the chemical material of the genes is the DNA (and not protein as it was erroneously believed). It means that the gene is a piece of DNA that encodes a gene product, usually a protein.

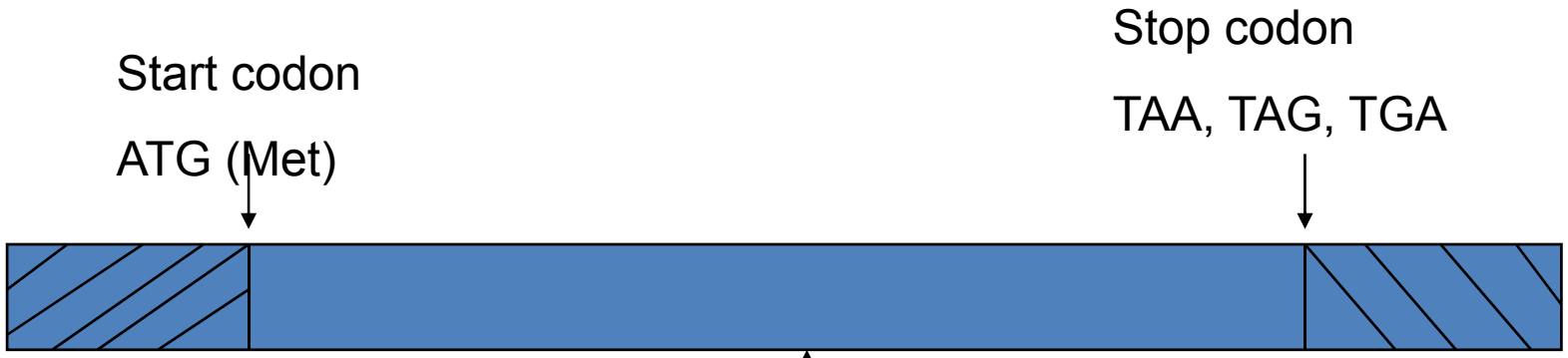
Present definition of the gene:

The gene is a segment of the DNA molecule that encodes the information required for the synthesis of a gene product (protein or RNA).

The term „protein-gene" usually refers to the well-defined coding region which encodes the amino acid sequence of a protein.

There are however regulatory sequences as well, that guide and control the gene expression (promoters, enhancers, operators, terminators, etc.). These sequences are also integral parts of a gene and therefore must be included in the definition.

The gene


Start codon
ATG (Met)
Stop codon
TAA, TAG, TGA
5’ noncoding region
promoter, enhancer, ribosome binding site
Coding region
ORF: open reading frame
3’ nonconing region
Polyadenylation signal
Transcriptional terminator

When we mention gene we usually mean gene that encodes for a protein.

There are however RNA genes too which determine the nucleotide sequence of an RNA molecule that will be not translated into protein.

Examples of such RNA molecules:

Ribosomal RNAs (rRNAs): The most intensely transcribed genes in all cells (nucleolus).

Transfer RNAs (tRNA): These RNAs play a key role in protein synthesis on the ribosomes (transcription).

Small nuclear RNAs (snRNAs): They are involved in the processing of mRNAs (splicing).

Examples of RNA molecules cont.:

Small nucleolar RNAs (sno RNAs): Participate in processing of other RNA molecules, such as rRNA, tRNA, snRNA.

Micro RNAs (mRNAs): about 22-nucleotide-long RNA molecules that are generated from longer precursor RNA molecules. They are included in the regulation of the gene expression.

The RNA genes have quite different structure in the genome compared to the protein genes. That is why these RNA genes are not easy to find. Actually, the genes coding for the precursors of miRNAs have been discovered only recently.

In the DNA molecule for basis (nucleotides) encodes the information: A (adenine), G (guanine), C (cytosine), T (thymine).

In the RNA instead of thymine we can find U (uracil).

In the proteins there are twenty amino acids: alanine (Ala, A), asparagine (Asn, N), aspartate (Asp, D), arginine (Arg, R), cysteine (Cys, C), glutamine (Gln, Q), glutamate acid (Glu, E), glycine (Gly, G), histidine (His, H), isoleucine (Ile, I), leucine (Leu, L), lysine (Lys, K), methionine (Met, M), phenylalanine (Phe, F), proline (Pro, P), serine (Ser, S), threonine (Thr, T), tryptophan (Trp, W), tyrosine (Tyr, Y), valine (Val, V).

The nucleotide sequence of the DNA determines the nucleotide sequence of the RNA and the amino acid sequence of the protein.

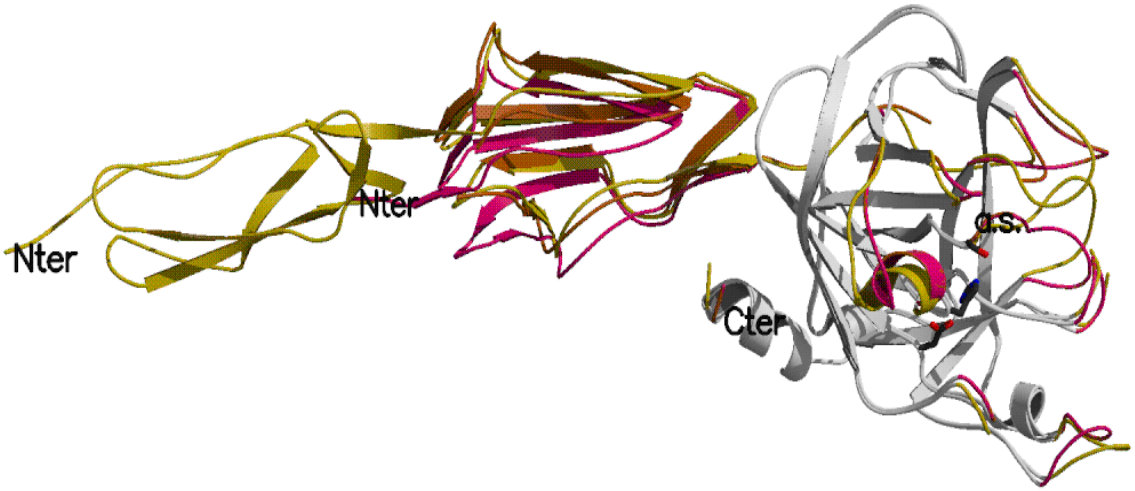
Three nucleotides (codon) corresponds to an amino acid in the protein (the genetic code).

The sequence of the amino acids in the protein determines the three dimensional structure of the protein.

The three dimensional structure of a protein is the prerequisite of the biological function.

The three dimensional structure of a protein is encoded in the amino acid sequence. We do not know the exact nature of the code. The nucleotide → amino acid code is straightforward (i.e. the genetic code). The translation of the nucleotide sequence into the protein sequence requires a sophisticated molecular apparatus (ribosomes, tRNAs, mRNA, associated proteins). In his famous experiment Christian Anfinsen proved that the amino acid sequence of a polypeptide chain contains all the information required to fold the chain into its native, three dimensional structure. A denatured polypeptide chain, under optimal conditions, can spontaneously refold into its correct three dimensional structure.

The three dimensional structure of a protein



This globular protein has three domains i.e. independent folding units.

The genomes of different living organisms can differ in size, structure, and information.

Size (kbp=1000bp, Mbp=10^6bp) / number of genes

ΦX-174 bacteriophage: 5.4 kbp / 10

Escherichia coli: 4.6 Mbp / 4377

Yeast (S. cerevisiae): 12.5 Mbp / 5770

Nematode worm (C. elegans): 100.3 Mbp / 20958

Plant (A. thaliana): 115.4 Mbp / 25498

Fruit fly (D. melanogaster): 128.3 Mbp / 13525

Human (H. sapiens): 3223 Mbp / ~23000

In eukaryote genomes it is more difficult to locate a gene than in the prokaryotes. RNA genes are even more difficult to find.

Gene finding (genome annotation) is one of the primary tasks for bioinformaticians.

C-value paradox:

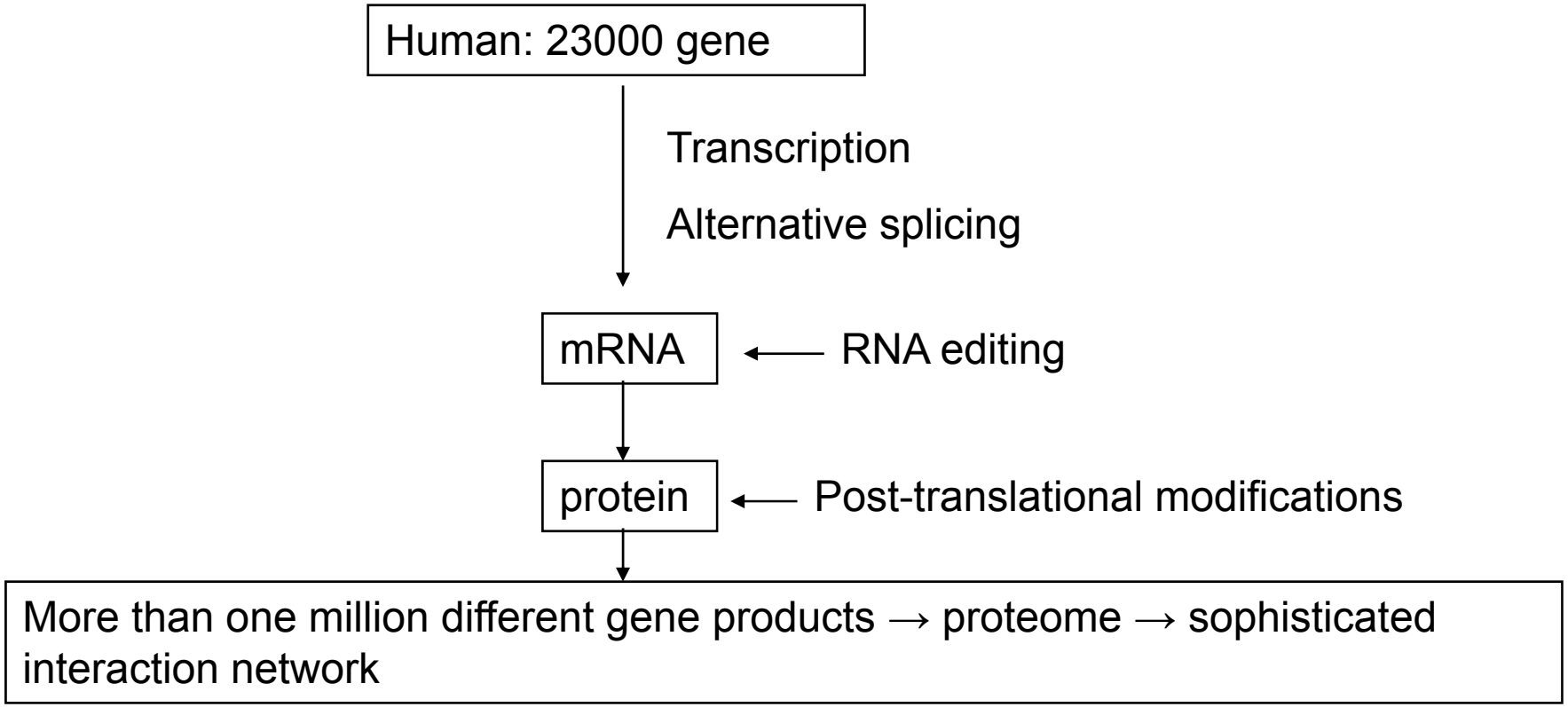
Genome size does not correlate with the complexity of a living organism. For example the single-celled amoeba has much larger genome than that of humans.

G-value paradox:

The number of the genes in an organism's genome does not correlate with the complexity of a living organism. For example plants have more genes than that of humans, and the nematode worm C. elegans has almost as many genes as that of humans.

Fundamental questions: What is biological complexity? Can we quantitate or measure it? Does evolution aspire to complexity?

The flow of information between the biological macromolecules is a source of diversity



3.) Structure of the genomes:

The toplogy and structure of the genomes of different organisms can differ significantly.

Prokaryotic genomes are closed circular double stranded DNA molecules. The protein coding genes are uninterrupted. There are no long intergenic regions.

Eukaryotic genomes consist of linear DNA molecules → chromosomes

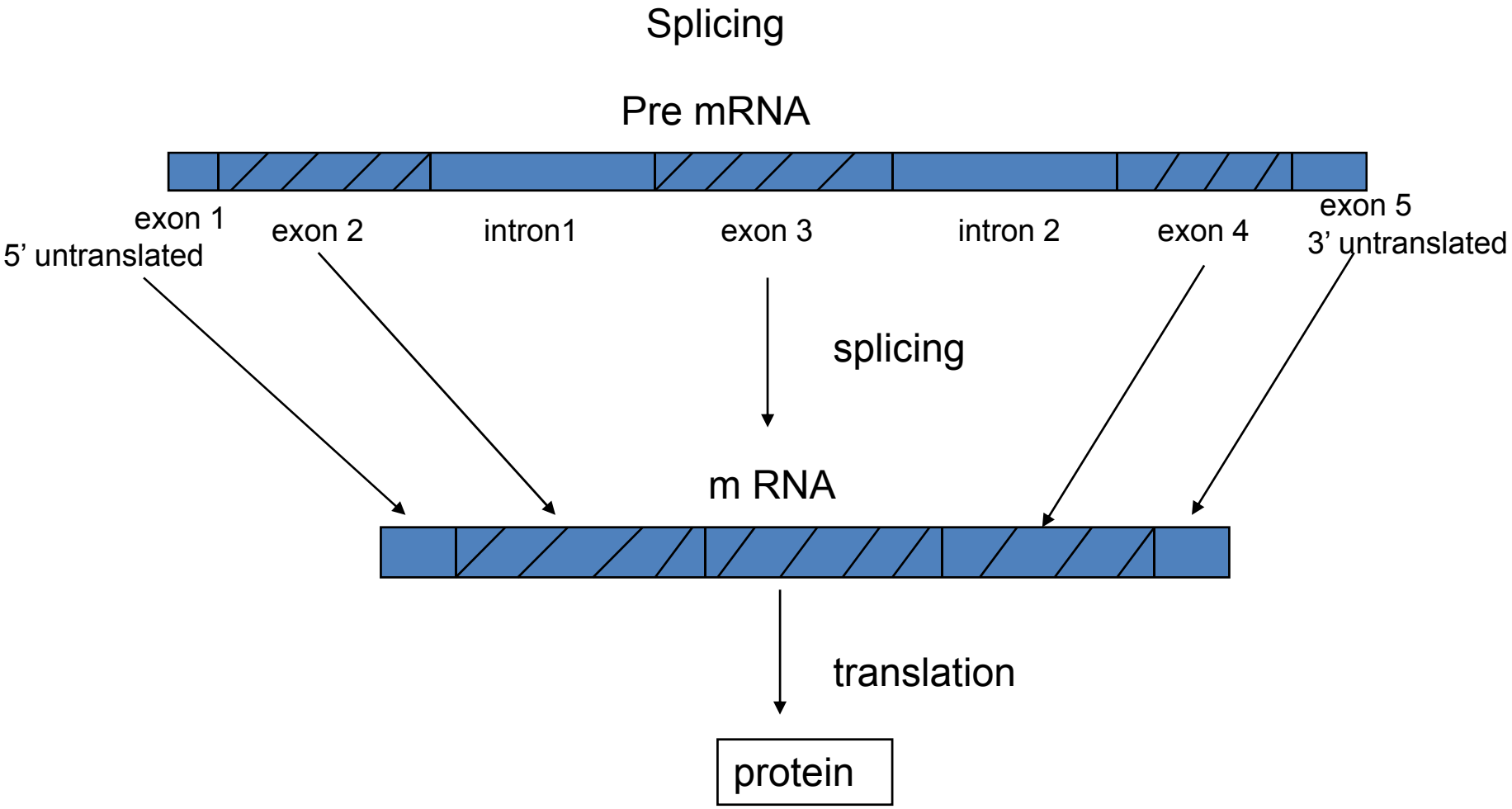
The protein coding region of most genes are not continuous but it is interrupted by noncoding sequences.

Exon: Segment of a eukaryotic gene that appears in the mRNS . The protein coding exons contain the codons (nucleotide triplets) that encode the amino acids of the polypeptide chain. There is a colinear relationship between the DNA sequence in the exons and the amino acid sequence in the protein. The exons are inrerrupted with introns.

Intron: Intervening sequence. The protein coding regions of the eukaryotic genes are interrupted by noncoding sequences called introns. Introns are transcribed but they are not present in the mature mRNA.

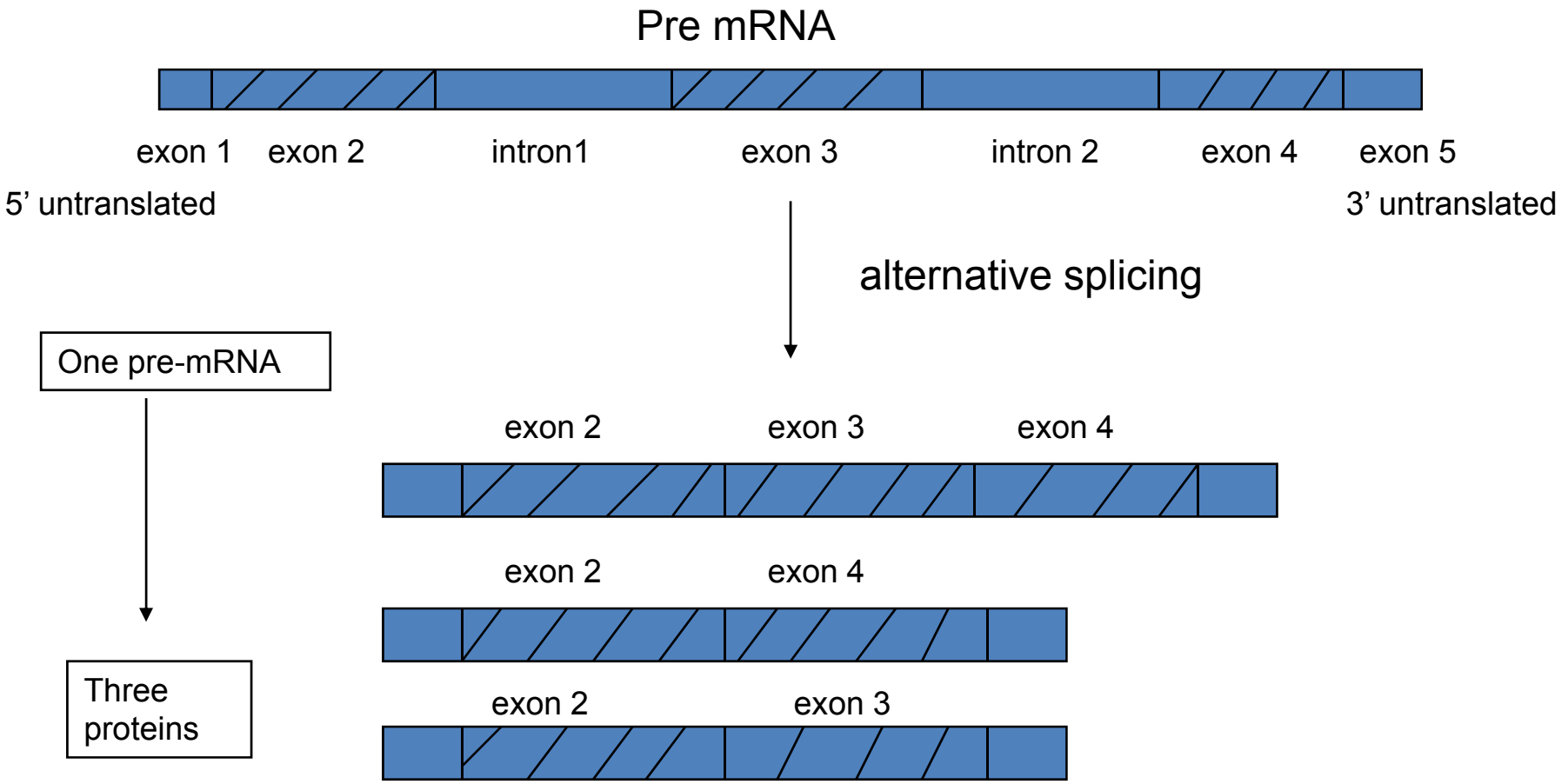
Introne can be much longer than exons.

Splicing: Immediately after transcription the nascent mRNA (primer transcript) contains the exons and the introns of the gene. During mRNA maturation the introns will be removed and the exons will be joined into a continuous piece of coding mRNA. This process is called splicing.


Splicing
Pre mRNA
exon 1
5' untranslated
exon 2
intron1
exon 3
intron 2
exon 4
exon 5
3' untranslated
splicing
m RNA
translation
protein

Alternative splicing: Different splicing reactions of the pre-mRNA of the same gene can result in different mRNAs that may be translated into different protein molecules. This process is typical among muntidomain protein, where alternative spilcing can add or remove domains. For example many proteins have membrane bound and free forms. The transmembrane domain which anchores the protein to the cell membrane can be added or removed at the mRNA level by alternative splicing. The alternative splicing is one mechanism which makes possible that a single gene encodes multiple protein molecules.

Alternative Splicing



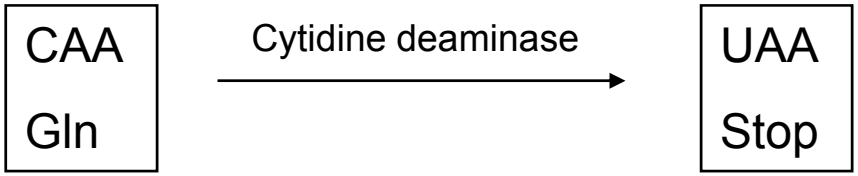
RNA editing: After transcription the information content of an RNA molecule can be changed by a process called RNA editing. RNA editing could mean base modification by chemical change, as well as nucleotide insertion. RNA editing has been observed in all major types of RNA (i.e. mRNA, tRNA, rRNA). If a mRNA molecule is modified by editing it will change the sequence of the amino acid in the polypeptide chain. In that case the primary structure of the protein cannot be predicted form the gene (DNA) sequence. RNA editing is another source of diversity that takes place at the RNA (post-transcriptional) level.

RNA editing of apoliporotein B (Apo B) mRNA

There is one Apo B gene in the genome, however there are two Apo B proteins: Apo B 100 (513 kDa) in the liver and Apo B 48 (250 kDa) in the intestine.

After translation a Stop codon is introduced in the middle of the mRNA and the translation will be terminated half-way at this point.

The Stop codon is created by the deamination of a cytidine.



Structure of the eukaryotic genome

The eukaryotic genome has distinct structural and functional elements:

1.) Genes and regulatory sequences:

Exons and introns

Regulation of transcription (promoters, enhancers, terminators)

Regulation of replication (origin of replication)

Regulation of translation

Sequences for recombination

Structure of the eukaryotic genome cont.

2.) Repetitive sequences

Highly repetitive sequences

Simple-sequence DNA

Satellite DNA

Moderately repetitive sequences

The precise role of the repetitive sequences is not yet understood. Some of them might have structural functions. The centromer of higher eukaryotes contains simple-sequence DNA. Telomeres also contain repetitive sequences.

Structure of a eukaryotic chromosome