Molecular diagnostics

Dr. István Balogh — Dr. János Kappelmayer

“Manifestation of Novel Social Challenges of the European Union in the Teaching Material of Medical Biotechnology Master’s Programmes at the University of Pécs and at the University of Debrecen”
Identification number: TÁMOP-4.1.2-08/1/A-2009-0011

University of Debrecen – Debrecen, 2011

The project is funded by the European Union and co-financed by the European Social Fund.

Manuscript completed: 17 November 2011
CONTENT

1. BIOLOGICAL INFORMATION 6
2. THE GENETIC CODE SYSTEM, TYPES OF MUTATIONS 11
3. TYPES OF MUTATIONS (CONTINUATION) 14
4. SPECIAL MUTATION CONSEQUENCES 18
5. GENETIC DISEASES CAUSED BY EXPANDABLE REPEATS - DYNAMIC MUTATIONS 22
6. MENDELIAN INHERITANCE 25
7. MULTIFACTORIAL DISEASES 30
8. EXAMPLES FOR MONOGENIC DISEASES 34
9. EXAMPLES FOR MONOGENIC DISEASES II 37
10. EXAMPLES FOR MONOGENIC DISEASES III 41
11. PHARMACOGENETICS 46
12. THE METHODOLOGY OF MOLECULAR DIAGNOSTIC PROCEDURES 51
13. THE METHODOLOGY OF MOLECULAR DIAGNOSTIC PROCEDURES II. 56
14. METHODOLOGY OF THE MOLECULAR DIAGNOSTIC PROCEDURES III 59
15. METHODOLOGY OF THE MOLECULAR DIAGNOSTIC PROCEDURES IV. 63
LIST OF FIGURES

Figure 1.1 Organization of the biological information from the perspective of genetics

Figure 1.2 From the gene to the protein

Figure 1.3 Consensus sequences of the exon intron boundary

Figure 2.1. The genetic code

Figure 2.2. Missense mutation

Figure 2.3. Nonsense mutation

Figure 3.1. Insertion

Figure 3.2. Duplication

Figure 3.3. Deletion

Figure 3.4. Frameshift mutation

Figure 4.1. Nonsense-mediated mRNA decay (NMD)

Figure 4.2. Pathogenic silent mutation

Figure 4.3. The use of the codons in humans

Figure 5.1. Genetic diseases that are caused by expandable repeats

Figure 5.2. Unusual DNA structures caused by expandable repeats

Figure 6.1. Autosomal recessive inheritance

Figure 6.2. Autosomal dominant inheritance

Figure 6.3. The family tree of a genetic disease inherited in X chromosome recessive way

Figure 6.4. The Age of founder mutations

Figure 7.1. Age-related macular degeneration (AMD)

Figure 7.2. Factors involved in the pathogenesis of Alzheimer disease

Figure 7.3. One type of familial Alzheimer disease. Genetics and the consequence of the mutation

Figure 8.1. Allelic disorders: one gene, three diseases

Figure 8.2. Linkage analysis for the identification of carrier status

Figure 9.1. CFTR protein

Figure 9.2. Monogenic disorders: cystic fibrosis (CF)

Figure 9.3. CF. Effect of the p.F508del mutation

Figure 10.1. Structure of the PKHD1 protein

Figure 10.2. NPC1 protein

Figure 10.3. Effects of the mutations: NPC1 gene
Figure 10.4. Blood coagulation: The protein C / protein S / Factor V system .................................................................44
Figure 10.5. Complementer tests for molecular genetic analysis: quantification of proteins by ELISA ...................................................45
Figure 11.1. Drug metabolism and excretion ........................................46
Figure 11.2. Genotype-phenotype associations in the case of CYP2D6 ....47
Figure 11.3. Human TPMT mutations ..................................................49
Figure 11.4. Pharmacogenetic aspects of vitamin-K cycle ..................50
Figure 12.1. From phenotype to genotype ............................................52
Figure 12.2. DNS isolation using silica microcolumns .......................53
Figure 12.3. PCR (polymerase chain reaction) ...................................55
Figure 13.1. Mutation screening methods: denaturing HPLC (dHPLC) ....57
Figure 13.2. Evaluation of the pathogenicity of mutations: protein truncation test (PTT) .................................................................58
Figure 14.1. Allele-specific PCR ............................................................60
Figure 14.2. Mutation detection using allele-specific oligonucleotide hybridization .................................................................61
Figure 14.3. Mutation detection using allele-specific oligonucleotide hybridization for the most common mutations causing cystic fibrosis. .....62
Figure 15.1. Mutation detection using oligonucleotide ligation assay ......63
Figure 15.2. Detecting mutation using hybridization probes - fluorescence resonance energy transfer ..................................................64
Figure 15.3. Detecting mutation using fluorescently labelled hybridization probes ........................................................................65
Figure 15.4. Detecting mutation using fluorescently labelled hybridization probes ........................................................................65
Figure 15.5. Multiplex Ligation-dependent Probe Amplification (MLPA) ..66
Figure 15.6. Different forms of DNA sequencing .................................67
1. Biological information

The main purpose of molecular diagnostical procedures is to detect qualitative and/or quantitative changes in the human genome. Molecular diagnostics in its wider sense involves all molecular biological methods that analyze the background of inborn errors, therefore, it contains in addition to the genetic and genomic approaches, the proteomic technologies as well. This chapter, however, deals with the narrower sense of molecular diagnostics that includes purely the analysis of nucleic acids.

The human genome consists of two main elements, DNA is present in the nucleus and in the mitochondria. Mitochondrial DNA is organized into one circular DNA molecule, which is 16.6 kilobases (kb) long. It is important from the molecular pathogenetic perspective that the mitochondrial DNA is inherited only through the mother. Mitochondrial DNA contains 13 protein-coding intronless genes, which are different from the majority of the nuclear genes, as these latter ones usually contain introns. The nuclear genome is 3.1 gigabases (Gb) long and it is organized into chromosomes. The human chromosome set contains 22 autosomes and two sex chromosomes, namely X and Y. There are approximately 20,000 protein-coding genes with a gene density of 1/120 kb. Human genes show extreme variability in their length and in their exon number. The average human gene consists of 10 exons. The highest number of exons in any human gene has been shown to be 363 (in the protein called titin, which is expressed in the muscle). The largest human gene is dystrophin (2400 kb). The encoded protein with the same name is also expressed in the muscle tissue. Mutations in the dystrophin gene lead to the Duchenne/Becker muscular dystrophy.

The functional expression of the genetic information coded in the DNA is regulated in many different ways (Figure 1.1).
Figure 1.1 Organization of the biological information from the perspective of genetics

The general direction of the flow of genetic information is DNA-RNA-protein. The sources of the original information is in the genomic or mitochondrial DNA. DNA is transcribed to RNA and the genetic information will finally be present in the expressed proteins (Figure 1.1a).

All levels are affected by complex regulatory processes, many of which are still unresolved in terms of details. Complexes that are formed between DNA and proteins define the active regions together with the epigenetic modifications of DNA. Transcriptional regulation can be observed in the different patterns of splicing, in the tissue-specific protein expression and in using small regulatory RNA molecules. Proteins, when translated, might be subject to extensive postranslational modifications (Figure 1.1b).

The average length of an exon is 300 base pairs. The main functional expression of the genetic information is protein synthesis. The genetic information coded in the genes is transcribed to RNA first, which is translated to protein. The primary transcript contains the entire sequence of the genes. During the maturation of the mRNA, the introns get spliced. The mRNA molecule will go through some modifications as described below:

- 5’ capping. mRNA molecule is modified in its 5’ end with a 7-methylguanosine. Main function of the 5’ cap is to protect mRNA from the 5’- 3’ exonuclease digestion to facilitate the transport of mRNA to cytoplasm to facilitate the splicing and binding of the ribosomal apparatus.
- Polyadenilation signal. The mRNA molecule is modified in its 3’ end with an approximately 200 nucleotide-long adenine tale. The site of addition is labelled by AAUAAA sequences motif. mRNA is cleaved approximately 15-30 nucleotide downstream from this signal and the adenines are added. Role of the polyadenilation signal are similar to that of the 5’ cap. It supports the cytoplasmic transport of mRNA, it stabilizes the mRNA and facilitates the recognition of ribosomal apparatus. A summary of the mRNA modifications is shown in Picture 1.2.

The importance of the sequence motives around the polyadenilation signal is highlighted by the fact that mutations occurring in this region might be pathogenic by interfering with the addition of the signal (prothrombin gene 20210A allele). Wild type allele in this case is the guanine at the nucleotide position 20210. When G to A mutation occurs, the result will be a more stable or better processed mRNA molecule, which will serve as a template for more effective protein synthesis. This increased amount of newly synthesized prothrombin protein will be secreted, the blood plasma level of prothrombin will be elevated as a molecular phenotypic consequence. The elevated plasma level of prothrombin increases the risk of venous thrombosis by 2- to 3-fold as compared to individuals not possessing the mutant allele. The importance of the above mentioned mutation is dual. It is not only a good example of the defect of the fine-tuning of the transcription regulation but it also marks a phenomenon whose prevalence rate might be as high as 2-4% in the general population.
Almost all protein-coding genes contain introns. Non-coding introns constitute the largest part of the gene, coding exons are normally much shorter. During the transcription process, the primary RNA transcript contains the introns and splicing will produce intronless mRNA. For the splicing to take place correctly, well-defined signals are necessary, which indicate the exon-intron boundary. Figure 1.3 shows a gene segment with one intron and 2 exons. The 5' end of the intron is the donor splicing site and 3' and is the acceptor site. Immediately preceding and following the exons, invariant GT and AG dinucleotides can be found in the intron (numbers above the letters show the relative occurrence of the given nucleotide). The adenine in the branch point sequences is also invariable. Y means either citosine or timine, N can be any nucleotide. Mutations that might occur in the invariant sites will have a fundamental effect on the process of splicing, as a result of which the proteins generated will often be decreased in amount or altered in their structure. In addition to the GT-AG introns, there are other types of introns, although they are very rare. The importance of the splicing signals is highlighted by the fact that almost all mutations affecting them are pathogenic.
Figure 1.3 Consensus sequences of the exon intron boundary
2. The genetic code system, types of mutations

The deleterious effect of the vast majority of pathogenic mutations are expressed via the translated protein encoded by the gene. The genetic code consists of 3 letters and each position (Figure 2.1.) can be occupied by any of the 4 nucleotides, therefore in theory there are 64 possibilities, which is more than enough to encode the 20 amino acids.

The genetic code is degenerated because every amino acid is encoded by more than one (normally 3) codon. Some amino acids (leucine, serine, arginine) are encoded by 6 codons, while others are significantly less well-represented. The third nucleotide of the codon can often wobble, which means that whichever nucleotide occupies the position, it will result in the same amino acid. The genetic code is almost universal but there are some exceptions. Chromosomal AGA and AGG codons that encode arginine, in the mitochondrion encode stop codons. ATA (isoleucine) is methionine in the mitochondrion, while TGA (stop) is tryptophane. TAA and TAG codons encode stop signal in both the mitochondrial and the nuclear genome.

![Genetic code table]

Figure 2.1. The genetic code

Mutations that affect the coding region of the gene are most frequently missense mutations. It means that an amino acid coding codon will be
replaced by another, but still amino acid coding codon. In Figure 2.2, CAT is replaced by CCT, which will result in the replacement of the originally coded histidine amino acid by proline. Approximately 70% of the single nucleotide replacements are missense mutations. The proportion of splicing mutations and mutations resulting in premature stop codons is less than 30%, while mutations in regulatory elements represent approximately 1%.

Figure 2.2. Missense mutation

The scope of the possible consequences of missense mutations is extremely wide. Very frequently, especially if the amino acid residue is solvent-exposed or if the amino acid is replaced by another amino acid with similar chemical characteristics, the effect is not crucial with respect to the structure and function of the protein. Conversely, the effect is much more severe when an amino acid that is involved in the maintenance of the structure of the protein or cysteine residue (which frequently forms disulphide bridge) is mutated. The change of one amino acid might completely destabilize a protein composed of a hundred or a thousand amino acids, resulting in intracellular degradation. The consequences of a missense mutation might be severe, for example a mutation in the active site of an enzyme. Unlike truncating and frameshift-causing mutations, predicting the effect of missense mutations is very difficult. Recombinant systems are often used for testing the effect of missense mutations. In this case, the coding mRNA is reverse transcribed into cDNA and that is cloned into a system which enables the recombinant protein production. This is usually done by cloning into an expression plasmid and the mutation is introduced using site-directed mutagenesis and the recombinant mutant protein is expressed. There can be many different kinds of expression systems. If the protein is
small, post-transcriptionally not modified and does not contain disulphide bridges, prokaryotic (most frequently E. coli-based) system can be the primary choice. In the case of proteins with more complex structure, eukaryotic (yeast or mammalian) systems can be used. Using these systems, it is possible to test the stability and structure-function relationships of the mutant proteins, which is why they are a valuable means of analysing the pathogenicity of the detected mutation.

Contrary to missense mutations, usually no experimental proof is needed if the genetic alteration results in the generation of premature stop codon. When an amino acid coding triplet is mutated into TGA, TAA or TAG codon, the result is nonsense mutation. The formation of stop codon usually results in truncated protein (if any protein is synthesized from such an RNA template). Premature stop codons are almost always pathogenic resulting in the encoded protein losing its function. When nonsense mutation occurs, often the mRNA will not be able to serve as a template for protein expression; instead, it is later degraded (see nonsense mediated mRNA decay). It means, that the cell possesses defense mechanisms at different levels (ie. mRNA and protein) in order to protect itself from a synthesis of a potentially deleterious mutant protein product.

Figure 2.3 shows a nonsense mutation, where the originally present CAG codon is replaced by a TAG codon, which will results in the finalization of the protein synthesis in this position.

**Figure 2.3. Nonsense mutation**

![Nonsense mutation diagram](image-url)
3. Types of mutations (continuation)

The previously described point mutations (missense, nonsense, splicing or regulatory) represent the majority of the pathogenic mutations. There are other types of mutations, where the originally present nucleotide number is changed. The size of these changes range from only one to a scale that can be detected by light microscope after the adequate staining of the chromosomes.

When one or more nucleotides are incorporated into a DNA sequence, the result is insertion. The effect of insertion largely depends on the number of the inserted nucleotides. If the number of the inserted nucleotides that are introduced into the codin region of a gene cannot be divided by three, the open reading frame is shifted. In this case, downstream of the mutation site the sequence of the encoded protein will be completely different and the resulting protein will be almost certainly unable to fulfill the function of the original. Insertions that cause frameshift usually result in the incorporation of the premature stop codon. In Figure 3.1., an adenine insertion can be seen, which causes frameshift, whereby the originally present histidines will be replaced by threonine and the following stretch of serine amino acids.

![Figure 3.1. Insertion](image)

Alterations that affect the genetic material are extremely diverse, ranging in their size from only one nucleotide (see above) to many thousands of

Identification number: TÁMOP-4.1.2-08/1/A-2009-0011
nucleotides. In some cases, even chromosome regions can be mutated, as it can be seen in Figure 3.2. The size of this mutation might reach the detection limit of light microscope or fluorescence in situ hybridization (FISH). The effects of this large-scale mutations, which might affect hundreds of genes, can be numerous. Depending on the affected region and/or the break points, they might have a role in the development of tumour, but the increased amount of the expressed proteins from the duplicated chromosome region might also be harmful.

Figure 3.2. Duplication

The opposite of the above mentioned insertion and duplication types of mutations is deletion. When one or more nucleotides are removed from a DNA sequence, the result is deletion. The effect of deletion is similar to the effect of insertion, in that it largely depends on the number of the inserted nucleotides. If the number of the deleted nucleotides introduced into the coding region of a gene cannot be divided by three, the open reading frame is shifted the same way as it is seen in the case of insertion. In this case, the downstream of the mutation site the sequence of the encoded protein will be completely different and the resulted protein will be unable to fulfill the function of the original. Deletions that cause frameshift usually result in the incorporation of the premature stop codon. In Figure 3.3, an adenine deletion can be seen, which causes frameshift and completely replaces the originally present histidine motif.
The open reading frame of the translation starts with the first ATG codon that encodes methionine. Insertion or deletion mutations, when the number of the affected nucleotides cannot be divided by three, will cause frameshift. A frameshift mutation is almost always pathogenic and results in premature stop codon. Such an mRNA molecule is usually degraded. It is not only small scale mutations that can cause a frameshift. In some cases, when entire exons are deleted or duplicated, the result can be a frameshift as well. A good example of that effect is one of the most prevalent monogenic diseases, the X-linked Duchenne/Becker type muscular dystrophy. In this case, the most prevalent cause of the disease are large deletions. The status of the open reading frame is an important prognostic factor, as mutations not affecting the open reading frame will result in Becker and mutations causing frameshift will result in Duchenne type of dystrophinopathy. Figure 3.4. shows the effect of a frameshift mutation, it does not show, however, whether the frameshift is caused by deletion or insertion.
Figure 3.4. Frameshift mutation
4. Special mutation consequences

Three codons of the 64 in the genetic code result in stop signal. These are the TAA, TAG and TGA codons. Therefore, terminating mutations are not rare, in fact, they account for 1/6th of all point mutations. Truncated proteins pose a significant danger to the cell. Their potential deleterious effect does not consist only in the fact that the resources are used for the generation of a useless end-product, but they are also prone to form aggregates in some cases, which might be poisonous for the cell. The cell has different protection mechanisms for the elimination of the potentially harmful truncated proteins. One of these mechanisms is nonsense-mediated mRNA decay (NMD), which affects the mRNA, in order to avoid the construction of the truncated protein. The mature mRNA has special protein complexes at the exact sites of exon junctions, called exon junction complexes (EJCs). As the ribosome moves in the mRNA, the EJCs are dislocated from the mRNA. However, in the case of an early (premature) stop codon, not all EJCs will be dislocated during the completion of the pioneer round of translation and the EJC that remained bound to the mRNA will trigger the NMD mechanism resulting in the degradation of the mRNA molecule (Figure 4.1.). NMD is not perfectly efficient, if the stop codon is in the last exon or too close to the EJC.

![Figure 4.1. Nonsense-mediated mRNA decay (NMD)](image)

As has been shown above, a detected molecular alteration can be tested to prove its harmful effect on different levels. If the mutation has been described in the literature as pathogenic alteration, this information is usually accepted without further investigation. The situation is different, however, in
the case of novel, potentially pathogenic mutations. If a point mutation introduces a premature stop codon, or causes frameshift, no further experimental work is needed. The case of the most frequent mutations, the missense mutations, is far more complicated. In addition to the experimental systems mentioned above, the physico-chemical consequence of the mutation (amino acid polarity, side chain composition, size) can be tested. Cis segregation of the detected mutation with the disease through generations is also an important indirect proof. Therefore, in order for a mutation to be qualified as pathogenic, it must cause some fundamental change in the structure and/or amount of the protein. This dogma, however, has been questioned recently.

Silent or synonymous mutation is a genetic alteration which does not affect the amino acid sequence of the protein. In the case of such a mutation, no phenotypic consequence is expected. In some cases, however, when a mutation affects a functionally important element, splicing defects might occur. In Figure 4.2, another mechanism is shown, where silent mutation might lead to gain-of-function of a protein. In this case, the kinetics of the translation will be impaired, which results in the changes of the conformation of the newly expressed protein. A c.3435C>T mutation has been described in the 26th exon of the MDR1 gene coding P-glycoprotein (Figure 4.2.). The originally coded isoleucine at the amino acid position 1145 will not be replaced by any other amino acid but the activity profile of the mutant P-glycoprotein will be somewhat different compared to the wild type. This difference might be attributed to the different tRNA that is needed for the mutant codon. The new tRNA can be considered to be a rare codon, which is why its availability is limited during the translation. The limitation will lead to different translation kinetics. The example above further complicates the problematics of potentially pathogenic mutations involving a single nucleotide, as the harmless nature of a synonymous mutation had been rendered unquestionable before the this description was formulated. However, there is no information as yet about whether such an effect is responsible for disease in any other gene.
Figure 4.2. Pathogenic silent mutation

As it has been shown, the genetic code consists of three letters and there are four possibilities within the codon, which allows for 64 possible combinations altogether. This number is more than enough to encode 20 amino acids. The third nucleotide of the codon often wobbles, which means that irrespective of the third nucleotide, the first two are responsible for the encoded amino acid. In Figure 4.3., the red numbers in the picture show the frequency of the given codon. This depends primarily on the intracellular availability of the given tRNA. The use of rare codons might affect the efficiency of the translation, which is a phenomenon that has been described in prokaryotic organisms. As regards humans, such a mutation has been shown to be present in the MDR1 gene (see above). In this case, the synonymous mutation causes change in the translation kinetics resulting in gain-of-function of the expressed P-glycoprotein.
Figure 4.3. The use of the codons in humans
5. Genetic diseases caused by expandable repeats - dynamic mutations

In addition to the classical monogenic genetic disorders that show Mendelian inheritance, there exists another group of genetic diseases, in which certain, meiotically unstable DNA repeats are the causing factors. A typical phenomenon to be observed in these disorders is anticipation, which means that the phenotype will be increasingly more severe in subsequent generations or it will appear at an increasingly younger age. These repeats form the mutation category called dynamic mutations. The number of repeats must normally reach a threshold value for the disease phenotype to appear. To date, approximately 30 such diseases are known (Figure 5.1.).

Figure 5.1. Genetic diseases that are caused by expandable repeats

Most of the cases show triplet expansions, but tetranucleotide, pentanucleotide and even dodecanucleotide expansions are possible as well. These unstable repeats might be present in different functionally important areas of the gene, like in the 5’ untranslated sequence region, in the introns, in the coding region, in the 3’ untranslated sequence region as well as in the promoter. The mechanism that leads to disease largely depends on where the expansion occurred in the gene.

- The repeat expansion in the 5’ untranslated sequence causes promoter methylation in the case of fragile X syndrome, and blocks the transcription of the gene.
- The CTG expansion in the case of dystrophia myotonica (myotonic dystrophy) in the 3’ end of the gene induces CUG-binding splicing
factors in the mRNA, which will interfere with the correct splicing processes in some independent genes.

- CAG expansion in the coding region will result in the expression of toxic protein product in Huntington’s disease.

Abbreviations (Figure 5.1.):

The number of the repeats shows extreme variability. In the FRAXA gene, the normal CGG repeat number is between 6 and 54. This can be expanded from 200 to 1000 in the case of unstable repeats. Myotonic dystrophy is a similar case. In healthy individuals, the repeat CTG number is between 5 and 37, which can be expanded to 50-10000 in patients. Because of the variable scale of the change in the number of repeats, the molecular testing of these disorders is very difficult. Different methodologies have been developed, such as special PCR assays and Southern blot.

Expansions occurring in the coding region of the affected gene are usually shorter than the ones located in the regulatory elements. In the case of Huntington’s disease, the number of original (normal) 6-34 CAG repeats can reach 36-100 in unstable cases. Repeat number in one of the spinocerebellar ataxia type (SCA6) shows only a small increase, from 4-17 in normal cases to 33 in affected individuals.

An important question to ask is what kind of molecular mechanisms play a role in the development of unstable repeats. It is thought that the instability might be a consequence of the disturbances in the DNA repair mechanisms and the recombination system. The repeats that are prone to instability form several unusual DNS structures (Figure 5.2.)
DNA repeats are capable of forming several unusual structures:
   a. imperfect hairpin structure formed by (CNG)n repeats
   b. (CGG)n repeats that form quadruplex-like structure
   c. Structures that are caused by (CTG)n (CAG)n repeats
   d. H-DNA and sticky DNA that are caused by (GAA)n (TTC)n repeats
   e. Structures that are caused by (ATTCT)n (AGAAT)n repeats

The repeats which do not form structures are much more stable genetically. The unusual DNA structures add a significant component to the instability of the DNA repeats, thus they contribute to a large extent to the development of the disease.
6. **Mendelian inheritance**

Almost all human diseases have genetic components. In the case of multifactorial disorders (see later), many environmental factors are added to the genetic background, while in the case of monogenic disorders, the genetic component can be exclusively responsible for the development of the disease symptoms. A few thousand monogenic diseases are known, and several more are becoming known continuously. With a few exceptions, such as mitochondrial diseases (showing maternal inheritance) and dynamic mutations (see above), monogenic disorders show classical Mendelian inheritance. Mendelian characters are located either on the autosomes or on the sex chromosomes. Depending on the mechanism underlying the genetic disease, the possible inheritance pattern can be autosomal recessive (AR), autosomal dominant (AD), X-linked recessive (XR), X-linked dominant (XD) or Y-linked. Approximately 65% of the monogenic disorders with severe phenotype show autosomal recessive inheritance, while the proportion of the inheritance patterns of autosomal dominant and the X-linked are 20% and 15% respectively.

In the case of genetic diseases which are inherited in autosomal recessive ways, it takes two wrong gene copies for the disease phenotype to develop. Those two mutations are inherited from the parents and they are not necessarily the same. If the child inherits the same mutation from both parents, the genotype will be a homozygous mutant. If the inherited mutations are different, the genotype will be compound heterozygous. The probability of children of heterozygous parents being heterozygous carriers is 50%, the chance of being wild type and homozygous (or compound heterozygous) genotype is 25 and 25% respectively. Therefore, in the case of a child who is not sick, the chance of being a carrier is 67%. Figure 6.1 shows the example of sickle cell anemia but similar picture can be dawn for several thousand monogenic diseases. In conclusion, in the case of autosomal recessive diseases, the affected child is born from unaffected, carrier parents. These disorders affect both sexes. It is important to note that inbreeding (e.g., in isolated populations) significantly increases the risk of the development of such a group of disorders.
In the case of diseases inherited in autosomal dominant way, only one wrong gene copy is enough to develop symptoms. Figure 6.2. shows the example of multiplex endocrine neoplasia type 2 (MEN-2). Both members of generation III has a 50% chance to inherit the disease. In those cases, the analysis of the genetic background, detection of the responsible mutation by some genetic method might help to establish the molecular genetic diagnosis before the onset of the symptoms, it will be possible, therefore, to implement some preventive measures in order to avoid the disease. In the case of autosomal dominant disease, both the child and one of the parents are affected (if the mutation is not \emph{de novo}). As the inheritance is autosomal, both sexes can be affected.
Disorders with the X chromosome might show either dominant or recessive inheritance. Figure 6.2. shows the inheritance of Duchenne muscular dystrophy (DMD). As the disease shows an X chromosome-linked recessive inheritance pattern, boys are affected in almost every case. There is no cure for DMD, so the molecular genetic diagnosis and carrier diagnosis are of great importance. Two generations can be traced back in the figure. The III/2 boy who is affected by the disease has a sister (III/1). In her case, due to the developments in the recent years in the molecular methods, it is possible to prove or exclude carriership. In the case of a heterozygous carrier female, it is possible to determine the genotype of the fetus in the case of pregnancy. The figure highlights an important phenomenon of the X-linked recessive diseases, namely, the family history of a previously affected male child in the mother's family.
Figure 6.3. The family tree of a genetic disease inherited in X chromosome recessive way

The situation is different in disorders showing X-linked dominant inheritance as they can affect both sexes. Female patients are involved in higher numbers, though with usually a milder phenotype. A female patient inherits the affected chromosome with a 50% probability, while all female children of an affected male will have the disease, none of his sons will (as they inherit his Y chromosome in those cases).

No known human disease is located on the Y chromosome. Apart from being responsible for the development of male sex, as it is absent in females, it cannot code any severe disease-causing gene. Y-chromosome microdeletions are frequently the underlying cause of male infertility. In addition to the sex-determining regions, there are so-called pseudoautosomal regions in both sex chromosomes as well.

Understanding the above-mentioned Mendelian inheritance types seems relatively easy. However, in practice, many factors might make the picture more complicated. If the case of a penetrance is not 100%, it might happen that the carrier does not show the phenotype associated with the disease. Mutations arising *de novo* result in a phenotype that has not been present in the family before. The situation can be further complicated by intrauterine fetal loss in very severe diseases, inbreeding or imprinting as well.

There are fundamental differences between the inheritance of nuclear and mitochondrial diseases. A special phenomenon of the latter is the maternal inheritance, as the sperm cells do not contribute mitochondria to the zygote. Another important difference is that in the case of mitochondrial diseases the heteroplasmy might be present, as the egg cell has more mitochondria. This
means that the even with the same mutation, the phenotypic expression of a mitochondrial disease can vary considerably.

A pathogenic mutation that affects an amino acid position can be formed because of the instability of the genetic material in the genomic region, but there are many examples of mutations that occurred a single time in the course of human evolution. This latter case is called founder mutation. The mutation might disappear in the following generations if evolutionally disadvantageous, or its frequency can be maintained or even increased if advantageous. The mutation site and genetic markers in the regions in close proximity might help us to determine the age of the founder mutation. The smaller the unit with linkage, the more recombinations occurred in the chromosome, so the older the mutation is (Figure 6.4.). A severe, rather common, but usually underdiagnosed iron overload disease, haemochromatosis, provides a good example of the founder effect. One of the most common causative mutations, HFE C282Y, came into being approximately 1500 years ago. Its prevalence nowadays is 4.5%. In this case, the selectionarily advantageous effect of the mutation could be more efficient iron absorption/storage in those periods of human history, when insufficient rather than balanced diet was the norm. This of course means that in our age, when such insufficient, iron-deficient diet no longer persists in Europe, the previously beneficiary mutation becomes harmful by establishing iron overload in the different tissues and organs, mainly in the liver causing irreversible damage through decades.

![Figure 6.4. The Age of founder mutations](image-url)
7. Multifactorial diseases

Monogenic diseases fall into the very heterogeneous group of rare diseases. By contrast, multifactorial diseases affect a large number of individuals. In almost all cases, multifactorial diseases have genetic components, but the relationship between the genotype and the phenotype is not a direct one. It means that the genetic factor does not cause the disease directly, but rather it contributes to its development together with the environmental factors (like diet, lifestyle, etc.). Although they have no significant role to play in the decision-making processes of the healthcare system and in therapy, defining these genetic factors are of great importance because the increasing knowledge about the pathogenetic routes might prove promising in terms of planning and designing future interventions.

The molecular genetic testing of the genetic risk factors of multifactorial disorders can be performed within the framework of either case-control or prospective studies. Methodologies in detecting the most common polymorphisms are highly developed, genome-wide association studies are frequently performed in order to define novel, previously unknown risk factors. Two different multifactorial diseases with different interesting aspects are detailed below.

Age-related macular degeneration (AMD) is a typical multifactorial disease. In addition to the numerous environmental factors (smoking, long-term sun exposure) that play a significant role in the development of the disease, AMD has a strong genetic predisposition. Those established risk SNPs are located in genes encoding proteins with unknown functions (for example LOC387715) and in a gene that is involved in the complement system (complement factor H (CFH) Y402H polymorphism). The disease that affects the region responsible for the detailed central vision in the eye is the most prevalent cause of blindness in the population above the age of 65 in the developed countries. The left-hand side of Figure 7.1. shows the more benign dry form of AMD with the typical lipid-rich deposits so called drusens. The wet form shown in the right-hand side of the figure is much more severe. Pathogenesis of the wet AMD involves neovascularization, therefore blocking the vascular endothelial growth factor effect using monoclonal antibodies is of great therapeutic importance.
Alzheimer disease is the most common cause of dementia, accounting for 2/3 of the cases. It is a progressive neurological disease with irreversible loss of neurons. Pathologically, Alzheimer disease shows loss of neurons and deposition of extracellular plaques with beta-amyloid peptides. These peptides are the result of a proteolytic cleavage of the beta-amyloid precursor protein. The pathogenesis of Alzheimer disease involves important genetic factors. The left-hand side of Figure 7.2. shows the genes and syndromes with proven association. However, mutations in the amyloid precursor protein gene (APP, chromosome 21) or in presenilin 1, 2 genes (PS1, PS2 in chromosomes 14, 1, respectively) can be detected in less than 2% of all cases, their discoveries are important milestones in acquiring more detailed information in understanding the pathomechanism. In the case of Down syndrome patients it is known that they have a life-long continuous mild production of beta-amyloid. Other genes labelled with a question mark highlight the fact that there are still unknown genetic factors to be discovered. In general, the familiar causes that are shown on the left-hand side of the figure lead to an early appearance of the disease, already at the age of 40-50. In the case of the nongenetic, environmental factors that are shown on the right-hand side of the figure, their pathogenetic role is not proved yet. In the general population, the most probable way of
pathogenesis is shown in the middle, where both environmental factors together with genetic predisposition (for example epsilon 4 allele of apolipoprotein E gene) and the physiological ageing process lead to the development of the disease. Alzheimer disease has a prevalence of 1% in the age group 65-69 and 40-50% above 95.

Figure 7.2. Factors involved in the pathogenesis of Alzheimer disease

Under normal circumstances, alpha, beta and gamma secretase enzymes cleave the beta-amyloid precursor protein (Figure 7.3. left panel). When the Ala692Gly mutation occurs in the beta-amyloid precursor protein gene, the consequence will be interference with the cleavage site as the 692nd position is close to the alpha secretase cleavage site (figure middle panel). Mutations Val717Gly, Val717Ile, Val717Phe interfere with the cleavage site used by the gamma secretase, which is the presenilin 1 itself or an essential cofactor of it (figure right panel). The consequence of the pathogenic mutations will be that instead of the 40 kD cleavage fragment, the 42 kD toxic beta-amyloid peptide will be the dominant cleavage product (red arrows show the dominant cleavage direction). Discoveries of the pathogenesis and the molecular genetic background make directed drug development possible.
Figure 7.3. One type of familial Alzheimer disease. Genetics and the consequence of the mutation
8. Examples for monogenic diseases

Duchenne/Becker muscular dystrophy (DMD-BMD) shows X chromosome-linked recessive inheritance. The disease is characterized by progressive muscle weakness and it is caused by mutations in one of the largest human gene, dystrophin. Dystrophin gene is 100-times bigger than an average human gene, its size is half of the entire E. coli genome. It is larger than any chromosome in the yeast. Its transcription takes 16 hours and provides the first mammalian example of the co-transcriptional splicing. It spans 2400 kb in the chromosome X and contains 79 exons. The encoded protein with the same name contains 3685 amino acids, its molecular mass is 427 kDa. The major function of the dystrophin is the mechanical reinforcement of the sarcolemma, by creating a physical link between the intracellular contractile elements and the components of the extracellular matrix. It binds with its N-terminal to actin and has contact with the dystroglycans. Two allelic diseases, Duchenne and Becker, are caused by different mutations in the dystrophin gene. These mutations in 2/3 of the cases are large deletions affecting one or more exons. Mutations that do not lead to the complete loss of dystrophin protein cause the much more benign Becker muscular dystrophy, which was described as an independent clinical entity before the identification of their common molecular background. The status of the open reading frame is an important prognostic factor. If the open reading frame is not changed, the muscular dystrophy will be Becker-type. In such a case, although the protein function is impaired, it is not completely abolished.

One third of the Duchenne/Becker muscular dystrophy cases are the consequence of novel mutations. The proportion is similar for point mutations and for gene segment duplications.

The laboratory diagnosis of Duchenne/Becker muscular dystrophy can be performed at different levels. Serum creatin kinase activity (released from the necrotizing muscle cells) is elevated in the patients, this overlaps, however, with the normal range in carriers. In the case of clinical suspicion, it is possible to analyze muscle biopsy for the presence of dystrophin protein. It can be done either by immunohistochemistry or SDS-PAGE followed by immunoblotting with a mono- or polyclonal antibody. The latter might help in determining the size (i.e. the level of truncation) of the mutant dystrophin and is able to give a “yes/no” answer as well (dystrophin present or absent).

Molecular genetic testing involves special PCR assays that are able to detect the presence or absence of the exons located at the deletion hot spots (17 exons and the promoter can be tested this way). These tests, however, cannot be used for carrier testing. To do so, another special amplification assay can be utilized, the multiplex ligation dependent probe amplification (MLPA, see later).

When it was determined that these two disorders are caused by mutations in a single gene, the clinical category of dystrophinopathy was created. It was soon realized that a third disorder, the X-linked cardiomyopathy also belongs to this group. This disease is caused by specific mutations that occur
in the heart specific promoter of the dystrophin gene. The consequence of these mutations will be the absence of dystrophin in the heart tissue while the skeletal muscle will have a normal or close to normal dystrophin level (Figure 8.1.). Therefore, the discovery, cloning and characterization of the dystrophin gene helped to set up the molecular associations in three different clinical diseases and in the development of the molecular genetic diagnosis.

![Diagram of dystrophin expression]

**Figure 8.1. Allelic disorders: one gene, three diseases**

Duchenne/Becker muscular dystrophy provides a good example of the presentation of the indirect molecular testing as well. Linkage analysis can be performed using intragenic or adjacent polymorphic markers when samples from members of more generations of the family are available and no information is available about the exact site of the disease-causing mutation. The testing shown on Figure 8.2. uses the a1, a2, b1, b2 polymorphic markers. In the case of this family, the a1-b1 marker combination defines the disease-causing X chromosome. In the third generation, the sisters of the affected male child can be tested for their possible mutation carrier status. III/2 member of the family is a carrier, therefore she can inherit the disease, while III/1 sister has not inherited the disease-causing chromosome.
Figure 8.2. Linkage analysis for the identification of carrier status
9. **Examples for monogenic diseases II**

One of the most prevalent severe monogenic inherited diseases is cystic fibrosis (CF). CF is caused by the abnormal function of the CFTR chloride ion channel. The CFTR protein consists of 1480 amino-acids with the domain organization shown in Figure 9.1. The largest part of the protein is located intracellularly. The two transmembrane domains (TM1, TM2, each composed of 6 transmembrane units) form the chloride channel. The nucleotide binding domains (NBD1, NBD2) have an important role in the opening of the channel. The carboxy terminal part of the protein (TRL, with threonine, arginine, leucine amino acids) makes contact with numerous other proteins and the cytoskeleton. These proteins significantly affect the function of the CFTR, the conductance, the localization and the mediation of other ion channels. The amino acid residue, which is affected by the most frequent CF-causing mutation p.F508del, is located in the NBD1.

![CFTR protein](image)

*Figure 9.1. CFTR protein*

CF is caused by loss-of-function mutations in the CFTR gene that result in abnormal function of the encoded chloride ion channel. CF is inherited in an autosomal recessive way. To date, more than 1600 different mutations have been described in the CFTR gene. One particular mutation, the deletion of
phenylalanine residue at position 506 is the most frequent mutation, being the underlying cause of approximately 70% of the cases. The other alterations show a large interethnic variability, which makes the search for population-specific mutations necessary. CF is the most common severe monogenic disease with a prevalence of 1:3000 and a carrier frequency (heterozygosity for one null allele) of 1:25. The classification of the molecular consequences of the different mutations can be seen in Figure 9.2. Class I mutations will result in the complete or almost complete abolishment of protein synthesis. These mutations usually cause frameshift, affect promoter or cause premature termination, and they are sometimes large deletions and insertions affecting several thousands of nucleotides. Class II mutations affect the folding of the expressed protein. The molecular consequence of these mutations will be the proteasomal degradation of the mutant CFTR proteins, without reaching the cell membrane. Class III mutations affect the regulation of CFTR necessary to its function. These mutations interfere with the nucleotide binding domains. Class IV mutations will cause impaired channel function. In this case, the channel either does not conduct or it is open for too short a time period. Some mutations might affect the turnover of the CFTR protein. The above-mentioned mutations will lead to the development of the disease when they are inherited either in homozygous or compound heterozygous form. Understanding the molecular background of the disease is very important as a novel class of drugs are being tested according to the mutational status (for example in the case of p.Gly551Asp), which means that personalized therapy might be possible in the near future.
Figure 9.2. Monogenic disorders: cystic fibrosis (CF)

The most frequent CF causing mutation, p.F508del is a deletion of three nucleotides in the CFTR gene with a consequence of the deletion of a phenylalanine amino-acid at the position 508. Figure 9.3. shows the maturation of the CFTR protein both in the case of wild genotype and of p.F508del. The immature B CFTR protein is located in the endoplasmic reticulum. In the case of the wild genotype a fraction of the newly synthesized CFTR to the status of stable B (although a small fraction is needed therefore approximately 70-80% will be directed to the proteasomal degradation way). In the case of p.F508del mutation all CFTR molecules are degraded. Wild type CFTR protein will acquire complex glycosilation in the Golgi and reaches the cell membrane. The middle part of the picture shows the immunoblotting detection of the CFTR proteins according to their maturation status. The detected molecular weights represent the difference in glycosylation, so it aptly represents the maturation status of the proteins.
Figure 9.3. CF. Effect of the p.F508del mutation
10. Examples for monogenic diseases III

The previously described Duchenne/Becker muscular dystrophy and cystic fibrosis are among the most common monogenic severe diseases. In the following, examples are shown for much more rare monogenic severe disorders, providing samples for the great variability for mutational spectrum, diverse clinical picture and methodological possibilities. The common feature of the diseases described is that, as the genes responsible for the diseases are cloned, prenatal molecular diagnostic procedures are applicable, especially, when the family-specific alteration is known (Factor V Leiden, being only a risk factor for venous thrombosis is excluded from this list).

1. Polycystic kidney and hepatic disease gene 1 (PKHD1).

Unlike the much more benign dominant counterpart, the autosomal recessive polycystic kidney disease (ARPKD) is a significant cause of neonatal morbidity and mortality accounting for a neonatal mortality rate of 25-35%. Prevalence of the disease is 1:20,000 with a carrier frequency of 1:70. The gene that is responsible for the ARPKD encodes a large protein with unknown function (fibrocystin or polyductin), see in Figure 10.1.

![Figure 10.1. Structure of the PKHD1 protein](image)

Identification number: 
TÁMOP-4.1.2-08/1/A-2009-0011
PKHD1 gene lacks any mutational hot spot, therefore, direct mutation analysis by gene sequencing is the recommended molecular genetic approach in the diagnosis of ARPKD.

2. Niemann Pick Type C disease (NPC).

The Niemann-Pick type C disease is rare monogenic disease that affects the intracellular transport of cholesterol. The disease affects quality of life to a great extent and is not curable. In 95% of the cases this disease is caused by mutations in the NPC1 gene. Figure 10.2. shows the predicted structure of the encoded NPC1 protein. The already known loss-of-function pathogenic mutations are labelled.

![NPC1 protein](image)

Figure 10.2. NPC1 protein

Figure 10.3. shows some amino acid residues affected by mutations causing Niemann-Pick type C disease. This domain is rich in cysteine amino acids, which are frequently involved in the formation of disulphide bridge. Double arrows show the already known pathogenic alterations. Amino acid residues in red are phylogenetically conserved, which means that they cannot be replaced by other amino acids.
3. Examples of inherited diseases in the blood coagulation system: multifactorial and monogenic disorders

During the process of the humoral way of blood coagulation, a fibrin network is formed, which closes the wound. A key component of this process is the activation of the final effector enzyme, thrombin (T, Figure 10.4.).

Figure legends:
1. Activation of the procoagulant factor V (FV). Activation is done by thrombin or by active factor X (FXa). The non-enzymatic FVa is a cofactor of FXa in the prothrombin - thrombin conversion.
2. The thrombin-thrombomodulin complex activates the natural anticoagulant protein C (PC).
3. Factor V has anticoagulant properties too.
4. The FXa / FVa complex in the activated membrane surface activates prothrombin in the presence of Ca^{2+}.
5. Activated protein C (APC) inactivates the procoagulant active factor V (FVa) in the presence of its cofactor, protein S by limited proteolysis that involves cleavage next to several arginine residues.
6. Procoagulant factor V can be inactivated by thrombin too.
7. Activated factor VIII is inactivated by APC/PS/FVac.
Mutations in the gene encoding FVIII cause haemophilia A. Loss-of-function mutations in the protein C or protein S genes result in severe thrombotic disease, while different mutations in the gene for factor V might result in severe bleeding as well as thrombotic tendency. Inherited factor V deficiency, which is a result of loss-of-function mutation in the factor V gene show autosomal recessive inheritance. Prevalence of factor V deficiency is 1:1,000,000. Contrary to this, when a specific mutation affects the primary cleavage/inactivation site for APC, the arginine at position 506, will result in the opposite effect, that is, increased risk of thrombosis (thrombophilia). The consequence of this mutation (the so-called factor V Leiden mutation) is that the mutant active, procoagulant factor V will stay longer in the circulation leading to increased risk of thrombosis. Heterozygous genotype increases the risk by 5-10-fold compared to wild type individuals, while homozygous mutant genotype increases the risk of venous thrombosis by 50-100-fold. The mutation being highly prevalent, it is also important at the level of public health. In Hungary, 1 out of 10 is heterozygous. Molecular testing of Factor V Leiden is one of the most commonly performed diagnostic assays in the developed world.

Figure 10.4. Blood coagulation: The protein C / protein S / Factor V system

For the final proof of a given inherited disease, molecular genetic assays are frequently complemented with protein-based tests, as it has been

Identification number: 44

TÁMOP-4.1.2-08/1/A-2009-0011
demonstrated in the case of Duchenne/Becker muscular dystrophy. Such an analysis is shown in Figure 10.5. The figure shows the result of antigen measurement in the case of a family with a factor V deficient member (generation IV, arrow). As has been shown, the autosomal recessive factor V deficiency is a rare disease with a prevalence of 1:1,000,000. Two mutations are needed to develop the symptoms, inherited in trans. According to the situation shown in the picture, on the father's side, the mutation can be traced back to the paternal grandmother (II. 2). Persons with heterozygous genotype will not have symptoms, as factor V levels around 50% are enough to maintain coagulation. The other disease-causing mutation (or more precisely, its consequence, the decreased factor V amount) can only be seen in the proband's mother (III. 2), which indicates its de novo generation. Antigen levels show that factor V deficiency in this family is a result of the decreased amount of factor V protein (CRM-, cross reactive material - ), meaning that the mutations interfere with the expression or stability of the mutant factor V.

Figure 10.5. Complementer tests for molecular genetic analysis: quantification of proteins by ELISA
11. Pharmacogenetics

One of the most quickly developing field of molecular testing is pharmacogenetics. In the process of metabolism, drugs will be more water soluble, therefore more accessible for renal excretion. During drug metabolism, sometimes toxic compounds are formed. In many cases, metabolism is responsible for the activation of the prodrug. This process can be divided into two different types of reaction. Type I reactions are oxdation, reduction and hydrolysis. Type II (conjugation) reactions include sulfation, methylation, glucuronidation, acetylation. Both reactions – whose names do not indicate the succession of the reactions – normally make the originally lipophilic compound more hydrophilic (Figure 11.1.) The genes coding the proteins responsible for these processes are usually highly polymorphic. This means that in some cases, the individual response after the administration of a specific drug can be attributed to the genetic background of the patient. It also means that in some cases, the knowledge of the patient’s genetic status makes individualized therapy possible. Individualized therapy has two major goals: it might help not only in quickly establishing the correct dose of certain drug, but also in avoiding the dangerous, sometimes life-threatening side effects.

Some pharmacogenetic examples are shown below.

![Figure 11.1. Drug metabolism and excretion](image)

Identification number:
TÁMOP-4.1.2-08/1/A-2009-0011
1. CYP2D6.
CYP2D6 is involved in the metabolism of many drugs, therefore it is one of the main pharmacogenetic targets (Figure 11.2.). Its gene is located in the chromosome 22. Null alleles (mutations that lead to non-functioning protein product or no protein product at all), labelled with white boxes will lead to the poor metabolizer (PM) phenotype. 5-10% of the Caucasian population belongs to this group. When a normal dose of the drug is administered to these patients, some severe side-effect might be experienced as a consequence. There are mutations which decrease enzyme activity, even though they do not completely abolish it (dotted boxes). Patients (5-10% of the Caucasian population) with this genotype show intermediate (IM) phenotype. Side-effects are also expected, though to a lesser extent compared to the poor metabolizers. In the case of wild type alleles (black boxes) the resulting enzyme activity is normal. Those individuals (65-80% of the Caucasian population) are the extensive metabolizers (EM). The duplication and multiplication of the CYP2D6 gene might be present in 5-10% of the Caucasian population, resulting in ultrarapid metabolizer (UM) phenotype. The administration of the normal dose of the drug is completely ineffective in those patients. The right-hand side of the picture shows the plasma concentration of the drug and its therapeutic range.

![Figure 11.2. Genotype-phenotype associations in the case of CYP2D6](image-url)

**Figure 11.2. Genotype-phenotype associations in the case of CYP2D6**
2. Thiopurine methyltransferase (TPMT) polymorphisms.

Thiopurine methyltransferase (TPMT) together with xanthine oxidase are responsible for the degradation of the purine analogue drugs, including 6-thioguanine, 6-mercaptopurine and azathioprin. There are significant interindividual and ethnic differences between the enzyme activities. It is the polymorphism of the TPMT gene that is responsible for these differences (Figure 11.3.). The TPMT gene is located to chromosome 6 (6p22.3). It consists of 10 exons and 9 introns. Numerous different alleles are known, which differ from each other only in a few nucleotides. The most common (wild type) allele is TPMT*1. The presence of any other allele results in decreased enzyme activity, both in heterozygous, and in homozygous form. The prevalence of heterozygosity is 11% among Caucasians. 1 in 300 is homozygous (phenotypically). The most common known alleles are TPMT*3A, 3B, 3C, 3D, 2, 4, 5, 6, 7, of which TPMT*3A-D are the most prevalent.

Individuals with either low or intermediate enzyme activity (homozygous or heterozygous genotypes, respectively) require much less than the standard dose of the above-mentioned drugs. The administration of the normal dose of the drug might result, especially in the homozygotes, in life-threatening side effects, such as myelosuppression and pancytopenia. The different common TPMT alleles are the followings: TPMT*1 is the wild type. TPMT*2 allele is a G238C replacement in the exon 5, resulting in an alanine-proline amino acid substitution. In TPMT*3A two point mutations are present in one allele: G460A in exon 7 (effect: alanine-threonine replacement) and A719G in exon 10 (tyrosine-cysteine substitution). G460A alone is TPMT*3B allele, while A719G is TPMT*3C. TPMT*3D allele has G292T mutation (exon 5, with a glutamine-stop consequence) in addition to the two mutations present in TPMT*3A. In TPMT*4 allele, the boundary of intron 9 - exon 10 is affected, with a G>A nucleotide substitution. TPMT*7 allele is T681G mutation in exon 10, resulting in a histidine-glutamine amino acid replacement.
3. The pharmacogenetic aspects of the anticoagulant therapy.

The most commonly used oral anticoagulant worldwide is warfarin. Its administration is lifelong in some patients groups (e.g., for those with artificial heart valves). One of the sister drugs of warfarin is acenocoumarol (-syncumar), which is very commonly used in Hungary (the use of warfarin is app. 10-15%). Warfarin interferes with the vitamin-K cycle (Figure 11.4.). The Vitamin-K cycle is a critical process for the posttranslational modifications of some proteins (the majority of them are involved in blood coagulation). The vitamin-K cycle makes possible the generation of gamma-carboxy-glutamate residues in these proteins, which is necessary for the membrane binding. Such proteins are factor II (prothrombin), factor VII, factor IX, factor X, protein C and protein S. In the absence of correct gamma-carboxylation the membrane binding, and as a consequence the blood coagulation is severely impaired. One of the key enzymes of the vitamin-K cycle is VKOR (vitamin-K epoxide reductase). The main function of VKOR is the reduction of the epoxide form of vitamin-K. This is a two-step process, where first the kinon, then the hydrokinon is formed. Warfarin interferes with both steps, as it binds to the same site where vitamin-K does and this site cannot accommodate both at the same time. INR is used for monitoring the effect of warfarin.
Pharmacogenetics of the warfarin/syncumar type of anticoagulant drugs is important from the point of view of both individual and public health. Both drugs exist in a mixture of optical isomers, but there are substantial differences. In the case of warfarin, the S isomer is responsible for most part of the anticoagulant effect, while in syncumar, it is the R. The key enzyme of the metabolism of the S isomers is CYP2C9. It has two common variants with decreased enzyme activity (CYP2C9*2 and *3 alleles). Although the VKORC1 gene is highly polymorphic, testing one functional variant, namely -1639G>A is sufficient. A wide international collaboration has resulted in a warfarin dosing algorithm, which was based on - in addition to other factors - the pharmacogenetic data. This dosing formula is available as an Excel worksheet.

In conclusion, the genetic status of CYP2C9 will affect the availability of the active drug (warfarin) and the genotype of VKORC1 will affect the efficiency of the vitamin-K cycle.

![Figure 11.4. Pharmacogenetic aspects of vitamin-K cycle](image-url)
12. **The methodology of molecular diagnostic procedures**

The large and diverse field of molecular testing includes the previously described monogenic and multifactorial diseases, the pharmacogenetic testing, but the methodology is the same in molecular oncology (pathology) and in molecular microbiology as well.

In many genetic diseases, the clinical diagnosis supported by some laboratory diagnostic test might make the direction of the further testing obvious (Figure 12.1.). If the gene is unknown, but its physical sites are known, linkage analysis can be performed using intragenic markers or markers close to the tested gene. In these cases, analysis of the members of the family for some generations is unavoidable. In the absence of such samples, the storage of the sample (biobanking) is the only solution until the genetic background of the given disease is clarified. In the case of a certain clinical diagnosis or at least suspicion (II) knowing the gene helps to a great extent the diagnostic efforts and the establishment of the genotype/phenotype associations. In those cases, the goal is to define the molecular background of the disease. If the mutations are scattered throughout the entire gene (absence of mutational hot spot), mutation screening methods can be utilized. These methods are SSCP, heteroduplex analysis, denaturing HPLC, high-resolution melting analysis. If the mutation affects large genetic structures, the use of citogenetic methods are necessary. If the mutation site is known (mutation hot spot), mutation detection methods (allele specific PCR, allele specific oligonucleotide hybridization, PCR-restriction digestion, hybridization or hydrolysis probes, DNA sequencing) can be used. Those methods are usually used in the case of mutation testing in common multifactorial diseases when diagnostic or predictive testing is performed.
The most frequently used material of the molecular genetic diagnostic procedures is the genomic DNA. In the past, DNA isolation was based on either phenol-chloroform extraction or salting-out procedures. These methods, however, are time-consuming and involve the use of toxic materials, therefore, they are not ideal for a diagnostic laboratory.

The next generation of DNA isolation, the spin column purification method is based on the phenomenon that DNA/RNA binds selectively to silica in high salt concentration. Cell lysates (usually originated from EDTA- or citrate anticoagulated blood) are loaded to spin columns containing silica membrane followed by several washing steps with different buffers, and the pure DNA/RNA is eluted with low salt concentration buffer or sterile molecular biology grade water. This method is the fastest available, but it is expensive. It involves many centrifugation steps and the yield of purified DNA/RNA is relatively little. The isolated DNA can be used further if its concentration and purity are adequate. DNA yield is determined by measuring the concentration of DNA by absorbance at 260 nm (at this wavelength one OD unit is equivalent to 50 micrograms DNA). The peak absorbance of DNA is at 260 nm and that of proteins is at 280 nm, thus purity is determined by calculating the ratio of absorbance at 260 nm to absorbance at 280 nm (Abs260/Abs280). Pure DNA has an Abs260/280 ratio of 1.7-1.9. Ratio less than
1.7 indicates protein contamination which might interfere with the downstream applications. DNA can be stored at +4°C for months, at -20°C for years, and at -70°C for at least ten years without significant degradation.

Figure 12.2. DNS isolation using silica microcolumns

Since its development in 1985, the PCR has revolutionized the possibilities of detecting minor changes in DNA (Kary Mullis, the developer of the method, received the Nobel prize in 1993). PCR is an in vitro technique to amplify a well-defined DNA fragment. PCR uses the physical characteristics of DNA (i.e. ability for reversible de- and renaturation), and the heat stability of a special DNA polymerase enzyme (Taq polymerase). The reaction consists of cycles. Each cycle starts with denaturation at a high temperature (95 °C) to separate the strands of double-stranded DNA template. The second step is fast cooling to 50-65°C to hybridize the specific oligonucleotide probes to the template and followed by an incubation at 72°C to synthesize the new strands.

The components of standard PCR reaction:

Identification number: TÁMOP-4.1.2-08/1/A-2009-0011
1. Template DNA.
2. Oligonucleotide primers. The primers determine the starting point of DNA synthesis (because the polymerase enzyme is able to initiate DNA synthesis only from a primer) and define the piece of DNA which will be amplified. It means that one of the primers will be hybridized to the sense and the other to the antisense strand. The orientation of DNA synthesis is 5'-3', therefore only the fragment between the two primers will be amplified.
3. Thermostable DNA polymerase enzyme. Mainly DNA polymerase in the Thermus aquaticus bacteria (living in thermal springs) (Taq polymerase) in native or recombinant form.
4. dNTPs. Deoxy-nucleotides in triphosphate form (dATP, dCTP, dGTP, dTTP).
5. Buffer (provides the necessary ionic strength and concentrations. It contains MgCl₂ which is a cofactor of the Taq enzyme).

Several different types of PCR are used in genetic diagnosis:
1. Standard PCR: this method is very useful in detecting small variations in genetic material (point mutations, short deletions and insertions), e.g., for the detection of factor V Leiden mutation, prothrombin 20210A allele, haemochromatosis.
2. Multiplex PCR: this method involves parallel amplification of a different piece of gene (i.e. exons) followed by agarose gel electrophoresis. This method is very useful in certain types of genetic diseases, for example Duchenne muscular dystrophy, where the disease is caused mainly (in about 70% of the cases) by deletions of exons.
3. Long PCR: this method combines two enzymes: Taq polymerase and a special enzyme which has a proofreading activity to remove the incorporated nucleotides in a wrong place. With this method it is possible to amplify larger DNA fragments (maximum 40 kb). Long PCR is a useful tool, for example, in the molecular diagnosis in Friedreich ataxia.
4. Quantitative real-time PCR: this method is used to quantitate the copy number of a specific DNA/cDNA in the sample. In routine diagnostic procedures, the fluorescent detection is the most widely used. During the PCR amplification, the fluorescence intensity (originated from a dye intercalating to the double stranded DNA or from a dye-labelled hybridization probe specific to the amplifiable DNA) is measured in every cycle. After the completion of the PCR, the software identifies the cycle in which the PCR product fluorescence signal has just exceeded the background (crossing point, CT). If standards with known copy number of the analyzed DNA are used, the standard curve will provide information about the copy number of the sample. This method is used in the testing of the translocation t(9;22) and the measurement of the bcr/abl transcript number in the sample.
Figure 12.3. PCR (polymerase chain reaction)
13. **The methodology of molecular diagnostic procedures**

In diagnostics, the molecular situation will be the basis of the chosen methodology. Mutation screening methods are used when the location of the possible mutation is not known in the gene. These methods are based on the PCR and use the physical properties of the DNA. Previously mainly SSCP, single strand conformational polymorphism, DGGE, denaturing gradient gel electrophoresis, and HA, heteroduplex analysis were used.

When an SSCP analysis is performed, first the PCR reaction is carried out. After the completion of the PCR, the sample is heated then quickly cooled. The quick cooling step does not allow the complementsing DNA strands to find each other, instead, intramolecular stabilization will occur. The resulting fragments are then separated by non-denaturing electrophoresis and the difference in the conformation that based on the sequence can be detected by the difference in mobility. The theory behind the heteroduplex analysis is very similar, though, in that case the cooling step is slow, which is why the complementsing strands are able to find each other. In the course of DGGE, the denaturation is performed during a gradient gel electrophoresis.

Later, the development of the suitable instrumentation has made some automation possible (dHLPC, Figure 13.1.). This method is based on the presence of both possible alleles, wild type allele and mutant allele. The PCR is followed by denaturation and a rehybridization step. In the presence of a mutation, two types of complexes can be formed:
- homoduplexes, where the wild type sense strand is hybridized with its wild type antisense strand, and the same occurs with PCR products possessing a mutation,
- heteroduplexes, where the wild type sense strand is hybridized with the mutant antisense strand, and the mutant sense strand is hybridized with the wild type antisense strand.

After the formation of the above-mentioned complexes, the next step is running the samples on an HPLC machine under partially denaturing conditions. The mobility of the heteroduplexes and homoduplexes are different, therefore the number of peaks indicate the presence of mutation. This method, similarly to the other mutation screening methods, does not provide information about the nature and location of the mutation, but only shows its presence within the PCR product. Therefore, the samples showing aberrant mobility need to be analyzed further using other methods, like DNA sequencing in order to determine the consequence of the detected alteration (i.e., harmless polymorphism or synonymous mutation, or pathogenic that causes the disease).
There are situations, where no hot spot is known in the disease-causing gene, but it is established that the disease is mainly caused by truncating mutations rather than missense changes. In such a case, protein truncation test (PTT) might be help in the establishment of the molecular genetic background of the disease.

PTT can be used for the analysis of mutations which affect the splice site, causing nonsense or frameshift. It is based on in vitro transcription and translation. The starting material of PTT is usually mRNA or a large exon of genomic DNA. mRNA is reverse transcribed into cDNA. Using modified primers, start codon is incorporated into the sample and the in vitro transcription and translation are performed. The resulting protein products are separated using SDS-PAGE and then detected with a specific method (i.e., by using antibodies). The special nature of the method is its advantage and drawback at the same time: it is capable of the analysis of mutations that replace the original stop codon (either by truncating the expressed protein, or by generating longer product by mutation of the original stop codon) only, the most common missense mutations will not alter the length of the expressed protein. Lanes 3, 8, 11 are mutations leading to truncated protein, while lane 5 represents a mutation that has removed the original stop codon. As the in vitro system does not contain the relevant components...
of nonsense-mediated mRNA decay, PTT will not answer the question whether or not the mutated product will reach its protein level at all. PTT has not become a popular screening method because it is complicated and labor-intensive.

Figure 13.2. Evaluation of the pathogenicity of mutations: protein truncation test (PTT)
14. Methodology of the molecular diagnostic procedures III

It has been shown before that although unknown mutations can be investigated in some cases, most of the molecular testing is performed in order to find known gene variants, like mutational hot spots, founder mutations, molecular risk factors or mutations/polymorphisms of pharmacogenetic importance. Traditionally, the most widely used method for the detection of point mutations is the PCR followed by restriction digestion (PCR-RFLP, restriction fragment length polymorphism). As the bacterial restriction endonucleases recognize a very specific sequence motif, they are very useful for the detection of mutations affecting restriction sites. There are a few hundred such enzymes. During the testing, the last step is an electrophoresis on agarose or acrylamide to separate the fragments according to their size. The number and size distribution of the restriction fragments are the function of the genotype.

Allele-specific PCR method can be used, when the site of the mutation is known, as a mutation detection method. Allele-specific PCR is based on the fact, that Taq polymerase enzyme needs perfectly hybridized primer (on its 3' end) for its function. In the case of 3' incomplementarity, Taq enzyme is unable to use the oligonucleotide as primer. The system contains a common oligonucleotide (right side of Figure 14.1) and contains primers specific to the possible alleles (one is shown on the left-hand side of the picture). Allele-specific PCR reactions are separated physically. With careful optimization, allele-specific amplification is possible. In the upper part of the picture a perfect hybridization is shown resulting in PCR product. In the lower panel, no product is formed due to the 3' end incompatibility. Using several pairs of primers, multiplexing is possible. The method is capable of showing all possible genotypes, i.e., wild type, heterozygous, homozygous.

Allele-specific PCR has been popular under many names, like ASO-PCR, ARMS. There are commercially available mutation detection diagnostic kits that are based on this principle.
Although the allele specific PCR has significant advantages compared to PCR-RFLP as it can be multiplexed, it still requires electrophoretic separation. There have been many efforts to develop a methodology that does not involve electrophoresis during allele discrimination.

Allele specific oligonucleotide hybridization is a general method distinguishing between alleles that differ by even a single nucleotide substitution (Figure 14.2.). The method is based on the difference in the melting temperature between perfectly matched and mismatched probe-template hybrids. In this procedure the allele specific oligonucleotides (ASOs) are immobilized on a nitrocellulose or nylon membrane and hybridized with a labelled PCR product spanning the variant nucleotide site. The discrimination between the two alleles is based on the fact that in a specific hybridization temperature the perfectly matched hybrid is more stable than the mismatched one. After hybridization and washing the detection is mainly colorimetric: for example if the PCR product is labelled with biotin, the detection is based on the streptavidin-alkaline phosphatase conjugate (or horseradish peroxidase) enzyme reaction with a substrate. By using two ASOs, it is possible to determine each combination: wild type, heterozygous, homozygous. The advantage of this method that multiplexing is possible (at least 5-10 different mutations are detectable at the same time) and it does
not involve a gel electrophoretic step, though many washing steps are needed and high stringency conditions are required. There are commercially available mutation detection diagnostic kits that are based on this principle.

![Figure 14.2.Mutation detection using allele-specific oligonucleotide hybridization](image)

One of the most important use of the allele specific oligonucleotide hybridization is shown in Figure 14.3.

Mutations in the CFTR gene which cause cystic fibrosis (CF) are very heterogeneous. To date, more than 1600 CF-causing mutations have been described of which approximately 30 are more common. There are mutation testing panels to analyze the most common alterations and the available diagnostic kits include these mutations. The figure shows such an assay. The test system, which is highly multiplexed, is performed using two membrane strips. Mutations are represented both with their wild type and mutated alleles. The left-hand side of the picture shows the CFTR19 and right side shows the CFTR17+Tn kit. Using both strips, many different mutations can be tested simultaneously. Sample 1 has been shown to be wild type for the analyzed mutations while sample 2 is homozygous mutant for p.F508del, the far most common CF-causing mutation as judged by the respective signals: on the CFTR19 strip only the M.F508del allele-specific probe provides signal,
the W.F508del = W.I507del probe does not. Sample 3 has two mutations. M.G542X could be detected on the CFTR19 strip, and M.R117H gives a signal on the CFTR17+Tn strip. As in both cases the wild type probe also gave a signal, the genotype of the sample is compound heterozygous. Sample 4 was similar to sample 3. The detected alterations were p.S1251N and 394delTT. Using this – or similar – diagnostic approach, 80-95% of the pathogenic mutations causing CF can be detected, depending on the tested population.

Figure 14.3. Mutation detection using allele-specific oligonucleotide hybridization for the most common mutations causing cystic fibrosis.
15. Methodology of the molecular diagnostic procedures IV.

The oligonucleotide ligation assay has also been commercialized in some cases. Figure 15.1 describes the use of oligonucleotide ligation assay (OLA) for the detection of CFTR gene mutations. One mutation site is defined by two oligonucleotide probes. Hybridization of the common probe to the PCR product is independent of the presence of the given mutation. Allele-specific probes, however, hybridize only into the respective allele. After hybridization, ligation using a thermostable ligase is performed and the resulting ligated fluorescent probes are separated using electrophoresis. The method can be multiplexed using different dyes and probes with different length, which makes possible to the analysis of 20-30 different mutations at the same time.

Figure 15.1. Mutation detection using oligonucleotide ligation assay

Figure 15.2 shows the basics of the special real-time PCR assay, which is based on the fluorescence resonance energy transfer (FRET). On the top panel, there are two single stranded DNA probes labelled with donor and acceptor dyes together with the DNA sample to be analyzed. If the donor dye is excited with a light of an adequate wavelength, and the two dyes are
separated from each other (i.e. the two probes are in solution, not hybridized to the template DNA), it will emit light and gets back to its ground state (middle panel). If both probes are hybridized to the template DNA and light is introduced to the system, the donor dye will not emit at its emission wavelength, rather it will transfer its energy to the acceptor dye which then will be excited (FRET). It means that we can excite the donor dye and detect emission from the acceptor. The method is capable of a real-time following of the PCR reaction and it can be used for the detection of small-scale mutations if one of the probes is designed to cover the mutation site. In this case, after the completion of the PCR, slow heating is applied and the fluorescent signal is continuously monitored. The hybrid that is formed between the wild-type specific detection probe and the template DNA is more stable than in the case of mutation, therefore it will have higher melting temperature. This difference is indicated with a decrease in the fluorescent signal.

Figure 15.2. Detecting mutation using hybridization probes - fluorescence resonance energy transfer

Figures 15.3 and 15.4 show the localization of the detection (or sensor) probe and the other, anchor probe on the DNA template.
Figure 15.3. Detecting mutation using fluorescently labelled hybridization probes

The detection probe lies on the specific mutation site. In the case described in Figure 15.3., the sample has a mutation. Therefore, the detection probe which is complementer with the wide type sequence has a mismatch (fourth nucleotide position from the right). Fluorescent dyes are labelled with red color (3’ end of the anchor probe) and yellow color (5’ end of the detection probe). One of the dyes serves as donor and the other one as acceptor. Detection is based on the fluorescence resonance energy transfer (FRET). The counterpart of the picture is shown in Figure 15.4, where the detection probe that is complementer with the wild type sequence is hybridizing with wild type genomic DNA, therefore the complementarity is perfect.

Figure 15.4. Detecting mutation using fluorescently labelled hybridization probes

The multiplex ligation dependent probe amplification (MPLA) method is a recent development for the detection of small quantitative differences (i.e. gene duplications or deletions) in the genetic material even in heterozygous form which makes it useful for carrier detection in Duchenne/Becker muscular dystrophy (Figure 15.5.). Each probe set contains two specific oligonucleotides, one short and one longer and the assay can be multiplexed. The short probe contains a target-specific sequence (21-30 nucleotides) in its
3’ end and a 19 nucleotides long sequence in its 5’ end (which is identical to the PCR primer used downstream). The long probe, in addition to the target-specific sequence, contains a stuffer sequence and a second PCR primer sequence. The first step in the assay is a long hybridization period when the probe mix is hybridized to the genomic DNA. With the addition of a ligase enzyme, only the hybridized probes can be ligated, and the amount of the ligated probes is proportional to the amount of genomic DNA copy number. The next step is a PCR reaction with one primer pair. The stuffer sequences in the longer probes differ in a few nucleotides in length so the products of the multiplex PCR reaction can be easily separated using high resolution capillary electrophoresis. The result of the separation is an electropherogram in which the detected fluorescent signal-generated peak area is proportional to the original copy number present in genomic DNA.

![Diagram of MLPA process]

**Figure 15.5. Multiplex Ligation-dependent Probe Amplification (MLPA)**

The mutation scanning or detection methods described in detail in the previous lectures can be used for investigating the molecular context of a given DNA sample. Although they are very useful, the reference methodology for the analysis of small-scale mutations is DNA sequencing, i.e., the determination of the nucleotide order in a DNA strand. It has been developed by a British scientist, Frederick Sanger, who received one of his Nobel prizes
for the DNA sequencing method. Sanger sequencing, which was developed in
the late 1970’s, made the Human Genome Project possible. The left-hand
side panel of the Figure 15.6. shows the classic Sanger sequencing, which is
based on chain termination with dideoxy nucleotides. The method was rather
time-consuming and difficult to perform as it required bacterial cloning and
pouring of large acrylamide gels was necessary for the electrophoresis. In
addition, the detection was based on the incorporation of radioactivity into
the DNA. Later it become much easier with the development of some
automation and fluorescent detection methods, but it still requires gel (or
capillary) electrophoresis for the separation of the sequencing products. The
right-hand side panel of the Figure 15.6. shows the basics of a new era of
genetic testing, the next generation of sequencing. Electrophoretic separation
is not needed anymore in the case of next-generation DNA sequencing
methodologies, therefore, they are very high throughput but require highly
sophisticated bio-informatics.

It is generally accepted that the next-generation sequencing techniques
will become the most important step in the revolution of genetic analysis. It
is now clear that they will have an unprecedented impact on molecular
diagnostics. It will be soon possible to sequence the entire genome of an
individual relatively quickly and at an acceptable cost. The genomic way of
investigating human disorders is likely to change our way of thinking of
health and disease and the healthcare system for good.

Figure 15.6. Different forms of DNA sequencing