Želmíra Balážová • Zdenka Gálová
Martin Vivodík • Milan Chňapek
Katarína Ražná • Jana Libantová

Biotechnology in Plant Production II

SUA·FBFS
Faculty
of Biotechnology
and Food Science

Nitra **2023**

DOI: https://doi.org/10.15414/2023.9788055226897

Title: Biotechnology in Plant Production II

Authors: doc. Mgr. Želmíra Balážová, PhD. (AQ 2.87)
Institute of Biotechnology, FBFS, SUA in Nitra

prof. RNDr. Zdenka Gálová, CSc. (AQ 0.3)
Institute of Biotechnology, FBFS, SUA in Nitra

doc. Ing. Martin Vivodík, PhD. (AQ 0.8)
Institute of Biotechnology, FBFS, SUA in Nitra

doc. Ing. Milan Chňapek, PhD. (AQ 0.3)
Institute of Biotechnology, FBFS, SUA in Nitra

prof. Ing. Katarína Ražná, PhD. (AQ 1.2)
Institute of Plant and Environmental Sciences, FAFR, SUA in Nitra

Ing. Jana Libantová, CSc. (AQ 0.2)
Institute of Plant Genetics and Biotechnology, PSBC SAS in Nitra

Reviewers: Mgr. Martina Hudcovicová, PhD. – NAFC RIPP in Piešťany

doc. Ing. Jana Moravčíková, PhD. – DoB FNS USsCM in Trnava

Approved by the Rector of the Slovak University of Agriculture in Nitra on 6th December 2023 as an online textbook for the students of SUA in Nitra.
The publication of the textbook was supported by the project KEGA 027SPU-4/2021.

ISBN 978-80-552-2689-7

DOI: https://doi.org/10.15414/2023.9788055226897

Obsah

DOI: https://doi.org/10.15414/2023.9788055226897

1 INTRODUCTION TO THE STUDY OF THE BIOTECHNOLOGY IN PLANT PRODUCTION

The aim of this subject is the study and application of actual technological possibilities (biotechnological methods, procedures and techniques) aimed at the rationalization and intensification of plant production – production of economically important products.

Plant biotechnologies make it possible to address the following priority goals in a targeted manner:

1. Targeted creation of new plant genotypes using modern methods of cell and molecular breeding (tissue culture techniques and genetic engineering) and genetic engineering of plants is studied in *Biotechnology in Plant Production I*:
Creation of genotypes with pre-programmed characteristics:
a) Production performance (harvest)
b) Production quality (nutritional, technological)
c) Plant resistance against adverse biotic and abiotic environmental factors
d) Minimization of material and energy inputs during crop cultivation and plant cultivation
e) Creation of genetically modified plants (GMR) for special biotechnological applications (production of special bioproducts usable in the food, pharmaceutical and energy industries)

2. Development and application of predictive methods for the detection and identification of molecular markers (at the level of DNA and proteins) using the genetic markers of plants is studied in *Biotechnology in Plant Production II*:
a) Diagnosis of genes, metabolites and storage substances determining biologically and economically significant properties of plant products
b) Nutritional quality (content and fractional composition of proteins, AA – composition, digestibility, presence of antinutritional substances: antimetabolites, toxins, allergens, etc.)
c) Technological quality (content and fractional composition of stock proteins)
d) Other properties

2 GENETIC MARKERS OF PLANTS

The need to identify, differentiate and verify the plant genotypes arises from the practical requirements of breeding, testing and licensing of newly created varieties, their patent protection, multiplication in seed production, business relations in the purchase and sale of seeds, processing relations, protection of copyright, licensing and other rights. The individual genotypes (varieties, hybrids, populations) of agricultural crops differ from each other in morphological and agronomic characteristics and properties. However, the variability in these characteristics cannot be, and in fact is not, so great that to allow us to clearly differentiate the genotypes on its basis even within the spectrum of annually permitted varieties. The current and future trends in the field of identification of plant genotypes are aimed at analysing the levels of protein and DNA polymorphism.

On an international scale, the issues of copyright protection and seed testing are resolved by organizations such as UPOV (Union Internationale Pour La Protection Des Obtentions Vegetales) and ISTA (International Seed Testing Association), which have developed internationally recognized work procedures for identifying and differentiating the genotypes of some plant species and testing their genetic homogeneity. The issue of authorization and protection of new varieties is also addressed in Slovak legislation. When granting a Breeder's

DOI: https://doi.org/10.15414/2023.9788055226897

Certificate, the new genotype (variety) shall, according to the Act of the National Council of the Slovak Republic no. 22/1996 Coll., meet four basic criteria. The variety must be different, balanced, permanent and new.

The procedures for identifying and differentiating plant genotypes should be unambiguous, with a high distinguishing ability, fast enough methodically, manually simple and financially accessible. At the same time, they should have a high capacity, i.e. the ability to analyse several hundreds of samples simultaneously. The majority of agriculturally important crops are made up of sets of genetically different but very related variants within the individual species. They occur either in the form of domestically adapted ecotypes, or much more often in the form of cultivated varieties, actively produced by humans – breeders. There are several approaches to identifying the plant genotypes depending on the plant species. The mutual differentiation and differentiation of plant genotypes can be carried out at several qualitative levels, such as:

1. **Morphological-agronomic assessment:** a) Macro- and microanatomy
   b) Agronomic parameters
2. **Technological indicators:** a) Secondary metabolites
   b) Other chemical components
3. **Karyological observations:** a) Number of chromosomes
   b) Morphology of chromosomes
4. **Protein analysis:** a) Storage proteins
   b) Isozymes
   c) Protein sequence
   d) Study of antibodies – immunology
5. **DNA analysis:** a) Hybridization techniques – based on the principle of DNA hybridization
   b) Amplification techniques – based on the principle of PCR reaction
   c) Amplification-hybridization – a combination of the two previous techniques

Visual evaluation of agronomic and morphological traits and features of the genotype is still the currently used means for monitoring genetic variability. The disadvantages of this evaluation mainly include the lack of polymorphism of the monitored traits, influence of agroecological conditions on the genotype of the monitored plant, possibility of detecting these traits only at a certain developmental stage, and the subjective nature of the evaluation.

The karyological, immunochemical and technological methods of genotype differentiation are also marginal and insufficient, and are suitable only for determining the relevant species or specific genotypes.

The identification of genotypes (varieties, populations, lines, hybrids) is dependent on the plant's genome. In some plant species it is possible to sufficiently distinguish the genotypes by performing an analysis of stored proteins of mature grain, while in others it is sufficient to identify the DNA profiles. Sometimes, however, it is appropriate to use a combination of both types of molecular markers, which are based on the principle of gel electrophoresis of protein macromolecules or DNA fragments.

The endosperm storage proteins – gliadins and glutenins – are suitable protein markers in wheat and hordeins are suitable for the identification of the genotypes of barley. High polymorphism, high heritability of protein profiles and the possibility of genetic interpretation of electrophoretic protein profiles are among the prerequisites for the use of protein profiles as genetic markers. These conditions are almost completely fulfilled by the glutenin and gliadin markers in wheat and hordein markers in barley.

The basic difference between the protein analyses and DNA analyses is that protein polymorphism reveals the variability in the DNA coding sequences and DNA polymorphism reveals the variability in all regions of the genome.

DOI: https://doi.org/10.15414/2023.9788055226897

3 MOLECULAR MARKERS AND THEIR APPLICATIONS

Genetic marker – any gene whose expression leads to a phenotypic expression that can be easily detected. A genetic marker is mostly used to identify a cell, chromosome and individuals, such as transformants and their generations.

Genetic markers can be divided into two groups:
1. qualitative – discrete and mendelistically inherited, characterized as present or absent, interpretable by allelic models. They mark e.g. the flower colour, resistance to phytopathogens, etc.
2. quantitative (QTL) – they mark the traits encoded by a larger number of genes. They mark e.g. plant height, resistance to wilting, seed weight, resistance to drought, technological quality, etc.

A molecular marker is a DNA sequence that is easily detectable, different in various organisms, and its inheritance can be easily monitored. An ideal molecular marker is highly polymorphic, i.e. it must exist in different forms and these must be distinguishable.

It should meet the required characteristics:
- High degree of polymorphism
- Distribution of polymorphic markers throughout the genome
- Clarity and accuracy of analyses
- High reproducibility of the analysis within and between the test workplaces
- Inheritance of codominant traits
- Quick and manually undemanding analysis
- Low price, simple and wide availability of the method

Molecular markers have been used to identify genotypes, detect genetic diversity between organisms, identify hybrids, etc. Analyses and identifications of plant genomes based on molecular markers have produced a huge amount of information and databases. Extensive genetic maps have been created for several plants with the help of molecular markers in the plant system: wheat, barley, rice and many others.

Molecular markers have also found use in the study of plant evolution and phylogeny, which makes it possible to understand their development from the related wild genotypes by studying their genetic diversity and classify them into appropriate groups. Such studies not only provide information about their phylogenetic relationships, but also provide the possibility of identifying new and economically unexplored genes of high importance. In the past, breeders often created new varieties by crossing the original genotypes, but the supply of new genes was minimal. Therefore, if it is necessary to expand the genetic variability and increase the heterosis effect, it is necessary to select the genotypes of related wild species and use the genes that carry the desired characteristics from them. Tagged genes encoding important traits can be used to detect the presence of economically important traits in the new genotypes created in the breeding programs.

Systematic documentation and evaluation of plant genetic resources is based on molecular markers that capture variability at the DNA level, or of its protein products. These markers complement the morphological, agronomic and other characteristics necessary for the classification and differentiation of plant genotypes within a species. Their importance and contribution is great both for research and breeding, as well as for practical use. Molecular markers have found significant application mainly in the mapping of the plant genome and the

search for genetic markers. Regarding the methods to search for polymorphism, it is the search for individual types of markers – ranging from morphological markers to DNA markers.

The use of molecular markers and molecular and biochemical procedures and methods in working with the plant genetic resources enables:

• **Differentiation of genotypes:** For the differentiation of genotypes, highly specific primers amplifying the microsatellites containing the sections of DNA, i.e. microsatellites within the DNA coding sequence, are used. Their main advantage is their high reproducibility. Storage proteins are used as markers for the differentiation of genotypes in wheat.

• **Searching for new and less frequent alleles:** Collections of genetic resources offer a wide variability of traits that manifest themselves not only phenotypically, but also in the genotype itself. For example, new and less frequent alleles of the Glu-1 locus, encoding high molecular weight gluten subunits (HMW-GS), are also of interest in wheat. These proteins can be found by analyzing the genotypes of various types of wheat that were not previously grown in our country (e.g. subunits 13+16 and 14+15).

• **Detection of duplicates in collections:** Duplicate genotypes are mainly caused by an incorrect transcription of the genotype during registration. However, it is possible to identify the duplicates with genetic diagnostics and prove that two genotypes with the same name are different. For example, analyses of glutenins and gliadins are used for this purpose in wheat.

• **Functional genomics in work with plant genetic resources (PGR):** Molecular markers can be used to detect different alleles of a single locus, while their length variability can be related to their physiological function.

• **Biochemical and technological analyses:** Some economically important parameters of plant genotypes can also be analysed in the laboratory. Although they are genetically determined, the environment often and intensively changes the expression of the genes that encode them.

We use several types of molecular markers to work with PGR in the field of molecular biology and biochemistry. Protein markers are the storage protein fractions of mature seeds (prolamins and glutelins) and isozymes. From the large spectrum of DNA markers, these are mainly microsatellite markers (SSR), specific STS markers, EST-SSR markers and only occasionally other types of DNA markers based on the polymerase chain reaction.

3.1 Cereal proteins

Proteins belong to the basic compounds of living organisms, and they can be found in every cell. They are very complex substances with characteristic properties, structure and function. The classification of proteins is carried out from different points of view.
According to solubility in various solvents, proteins are divided into:

a) Albumins – soluble in water
b) Globulins – soluble in 0.1 – 0.5 mol.dm^{-3} NaCl a KCl
c) Prolamins – soluble in 60 – 80% ethanol
d) Glutelins – soluble in 0.2 – 2% NaOH
e) Histons – soluble in bases (NaOH)
f) Protamines – soluble in water and diluted salt solutions

DOI: https://doi.org/10.15414/2023.9788055226897

According to their functional significance, proteins are divided into:
a) protoplasmic proteins, which include:

- catalytic proteins – albumins and globulins. These proteins are enzymatically active, characterized by amylase and protease activity
- constitutional proteins – together with nucleic acids and lipids form the structures of the cytoplasm and the nucleus

b) storage proteins – these include prolamins and glutelins, which determine the technological quality of the grain

The biological value of proteins is part of their nutritional value. The biological value of proteins is understood as an agreement between the composition of amino acids of the given protein with the composition of amino acids of the proteins that are used to build the human or animal body. The human body cannot synthesize essential amino acids (lysine, histidine, arginine, etc.) by itself because it lacks the enzymes necessary for the synthesis of these amino acids and they must be supplied in food.

The lack of a particular essential amino acid in food or feed limits the use of all other amino acids, while the biological value and usability of proteins is determined by the very amino acid (limiting amino acid) that is present in the smallest amount. From the point of view of the amino acid composition of individual protein fractions, albumins are characterized by a high content of essential amino acids, such as lysine, threonine, methionine, isoleucine and tryptophan. Globulins are characterized by a high content of arginine and a significantly lower content of tryptophan and methionine. Prolamins are characteristic of their low content of lysine, arginine, histidine, tryptophan and a high content of glutamic acid and proline. Glutelins differ from prolamins in that they have a higher content of lysine, arginine, histidine and glycine.

The nutritional value of cereal proteins is low and is determined by a high proportion of protein fractions of the prolamin type, which are characterized by a low content of essential amino acids (lysine, methionine, arginine, etc.) and, on the other hand, a high proportion of non-essential amino acids (glutamic acid, proline). Other unfavorable features of the protein complex of cereal grains include the presence of proteins exhibiting anti-nutritive properties.

The biological role of storage proteins (prolamins and glutenins) in the cereal grain is to nourish the germinating embryo of the plant. They are heterogeneous in their chemical nature, consisting of several subunits. When dough is mixed, wheat gluten proteins (gliadins and glutenins) aggregate and form a structural complex – **gluten**, which plays a key role in the process of forming wheat dough and determines its baking properties. Based on its viscoelastic properties, the gluten complex has the ability to retain carbon dioxide, which is created in the process of fermentation of the dough by baker's yeast from fermentable sugars, which are partly present, but mainly released by amylolytic enzymes, thus enabling the maximum volume of baked goods to be obtained. From a nutritional point of view, they are incomplete because they show an unbalanced amino acid composition, a high content of glutamic acid and proline, and few essential amino acids, especially lysine.

Prolamins of wheat (gliadins), barley (hordeins), rye (secalins) and oats (avenins) are also known as initiators of the malabsorption syndrome of gluten-sensitive enteropathy, the so-called celiac sprue (coeliac disease). The incidence of this disease is constantly increasing and the fraction of alpha-gliadins formed by tetrapeptides has proven to be the most toxic.

Barley (*Hordeum* L.)

Barley (*Hordeum*) belongs to the Poaceae family. The barley grain endosperm prolamin proteins, which are also called hordeins, are synthesized on the endoplasmic reticulum of the developing endosperm cells and stored in the vacuolar systems as the so-called "protein bodies". Hordein is made up of monomeric and polymeric proteins and the polymeric proteins are stabilized by disulfide bonds. Hordein polypeptides are divided into two basic groups (B and C hordeins) and two smaller groups (D and gamma type hordeins). B hordeins and gamma-type hordeins are sulfur-rich prolamins, C prolamins are sulfur-poor, and D hordein is a prolamin with a high molecular weight. B hordeins make up 70 to 80% and C hordeins 10 to 20% of the total fraction, while D hordeins make up only 2 to 4% and the content of gamma hordeins has not been precisely determined. Genetic analyses indicate that B and C hordeins are encoded by the Hor1 and Hor2 loci and are located on the short arm of chromosome 5. D and gamma hordeins are encoded by the Hor3 locus located in the proximal region of the long arm of chromosome 5. The structural genes for hordeins have not yet been identified, but they probably correspond to the Hrd F (Hor5) locus, located near the Hor2 locus. This gene is located between Hor 1 and the centromere.

Rye (*Secale* L.)

Rye (*Secale*) belongs to the Poaceae family and Secale cereale L species. The high nutritional value of rye is associated with a higher proportion of albumin and globulin fractions, and the related higher average content of lysine in rye proteins compared to wheat. The majority protein fraction of rye grain is represented by secalins, which can be divided into three groups: high-molecular secalins with a molecular weight of 100 kDa, sulfur-poor ω-secalins with a molecular weight of 50 kDa, and sulfur-rich γ-secalins with a molecular weight of 40 or 75 kDa.

Oat (*Avena* L.)

Oats (*Avena* spp. L.) also belong to the Poaceae family. Naked oat (Avena nuda) is an agronomic variation of Avena sativa. Proteins are one of the basic components of oat grain and their content ranges from 12% to 16% in hulled oats and between 15% and 24% in bare oats. The embryo contains approximately 25% – 40% of proteins, 24% – 32 % can be found in the covering part of the grain, 18 % – 32 % in the bran and approximately 9 % – 17 % in the starchy endosperm. Oats contain proteins of very high quality with a significant presence of amino acids, especially lysine with an average of 4.2 % in the grain. The representation of individual protein fractions in oats is different compared to wheat, rye and barley. Prolamins are the most represented fraction of proteins in wheat, rye and barley, and globulins are the most represented fraction in oats.

Buckwheat (*Fagopyrum* Mill.)

Buckwheat belongs to the Polygonaceae family. Although it is not classified as a cereal from the point of view of botanical characteristics, it is included in the group of pseudocereals based on the chemical composition of the seed and similar methods of cultivation and use. Buckwheat proteins have a high biological value, but their digestibility is relatively low. The protein content of buckwheat ranges from 8 % to 19 % depending on the variety.

Compared to common cereals, buckwheat groats have an almost optimal representation of essential amino acids and, in particular, a high content of lysine, threonine, tryptophan and sulfur amino acids and a smaller proportion of non-essential glutamic acid, which is why buckwheat is an excellent supplement to common cereals. Buckwheat seeds do not contain prolamins, which are allergenic for celiacs, so products made from buckwheat flour are recommended for gluten-free diets. The content of prolamins in buckwheat ranges from 3.8 to 5.2 mg.100 g^{-1} in seeds.

3.1.1 Use of the HMW glutenin subunits in the identification of wheat genotypes

Storage proteins are the most widely used type of plant proteins for the purposes of identification and differentiation of genotypes. They are located in specialized plant tissues or organs. They can be found in them in sufficient quantities and are relatively easy to extract. Using basic electrophoretic techniques and their modification in polyacrylamide gels, these proteins can be separated, visualized and, in some cases, genetically interpreted in a simple and quick way. The procedures of electrophoretic identification, differentiation and characterization of plant genotypes have already been accepted by international organizations dealing with the protection of copyrights to varieties (UPOV – Union Internationale Pour la Protection des Obtentions Vegetables) and seed testing (ISTA – International Seed Testing Association).

The storage proteins of wheat (*Triticum*) grain are primarily represented by gluten, which is located in the endosperm of the grain. Gluten proteins (gliadins and glutenins) form around 80% of the total protein content of wheat grain. High heterogeneity, species and genotypic specificity, different physico-chemical properties as well as the fact that their biosynthesis takes place in the endosperm in the last phase of grain formation, predisposes them to fulfill the role of markers of economically significant traits and characteristics.

Classification of gluten:

a) Monomer units of gliadin (Figure 3.1):

- With a high content of sulfur amino acids: alpha-gliadins, beta-gliadins, gamma-gliadins
- With a low content of sulfur amino acids: omega-gliadins

b) Aggregated gluten:

- Low molecular weight components of gluten (LMW-GS), the so-called aggregated gliadins
- High molecular weight components of gluten (HMW-GS)

The separation of storage proteins by electrophoretic methods allows us to assess the nutritional and technological quality of wheat grain in terms of the content and component representation of HMW glutenin subunits, as well as the antinutritional properties of grain according to the representation of alpha-gliadin proteins.

Proteins are one of the first products of gene expression, and their composition can be used to predict the composition of genes and thereby collect the information about the genotype. One of the most important ways of obtaining the information about the variety is to study the protein polymorphism using electrophoretic and chromatographic methods. The principle of protein markers makes it possible to determine the origin of cultivated plants based on seed proteins, determine the structure of their genome and carry out their genomic analysis, accurately and quickly identify varieties, detect lines and mutants.

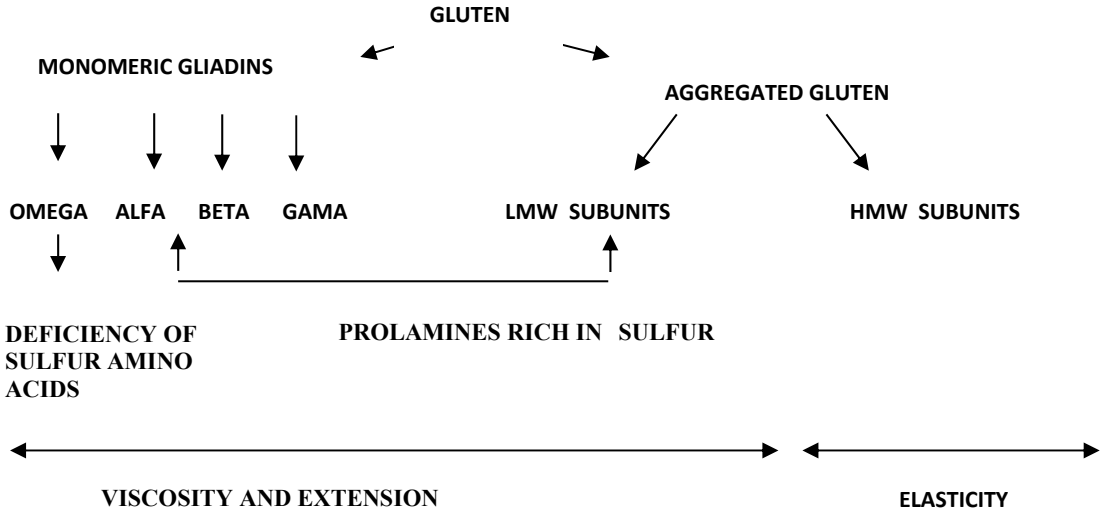


Figure 3.1 Fractional structure of gluten proteins

Each genome of hexaploid wheat (*Triticum aestivum* L.) is composed of seven triple pairs of chromosomes, while gliadins and glutenins are encoded by the genes of the following loci:

- The Glu-1 locus is located on the long arm of chromosomes 1A, 1B, 1D with genes encoding the HMW glutenin subunits.
- The Gli-1 and Glu-3 loci are located on the short arm of chromosomes 1A, 1B, 1D with genes encoding the LMW glutenin subunits, gamma – gliadins and omega – gliadins.
- The Gli-2 locus is located on the short arm of chromosomes 6A, 6B and 6D with genes encoding the alpha-gliadins and beta-gliadins.

Approximately 10 gliadin alleles have been described for each Gli-locus, which were identified based on their differential mobility during the polyacrylamide gel electrophoresis. For the LMW glutenin locus, 6 alleles (A-genome), 9 alleles (B-genome) and 5 alleles (D-genome) have been described. The nucleotide sequence of the genes encoding the HMW glutenin subunits has been described previously.

The **HMW glutenin subunits** are synthesized only in the endosperm of the developing wheat grain and their synthesis is controlled by the genes (Figure 3.2) located at three loci labelled as Glu-A1, Glu-B1 and Glu-D1.

With its HMW glutenin genes, Chromosome 1A determines the formation of five zones in the electrophoretic spectrum (A1-A5), while each zone corresponds to one individual glutenin subunit. The two weaker zones (A4 and A5) are indistinct and very difficult to detect. The molecular weight of subunits A1 and A3 ranges from 114,000 to 105,500 Daltons.

Chromosome 1B controls the synthesis of at least 12 distinct HMW glutenin subunits (B1-B12), with one or a pair of subunits encoded here. Zones B1 to B6 are characterized by slow mobility with a molecular weight of 92,000-102,000 Daltons. The fast-moving zones (B7-B12) show a molecular weight in the range of 84,000 - 91,000 Daltons.

Chromosome 1D accounts for at least six different zones of the HMW glutenin spectrum. The molecular weight of the slow-moving zones D1 to D3 ranges from 106,000 to 108,000 Daltons and the fast moving subunits D4 to D6 from 78,000 to 84,000 Daltons.

DOI: https://doi.org/10.15414/2023.9788055226897

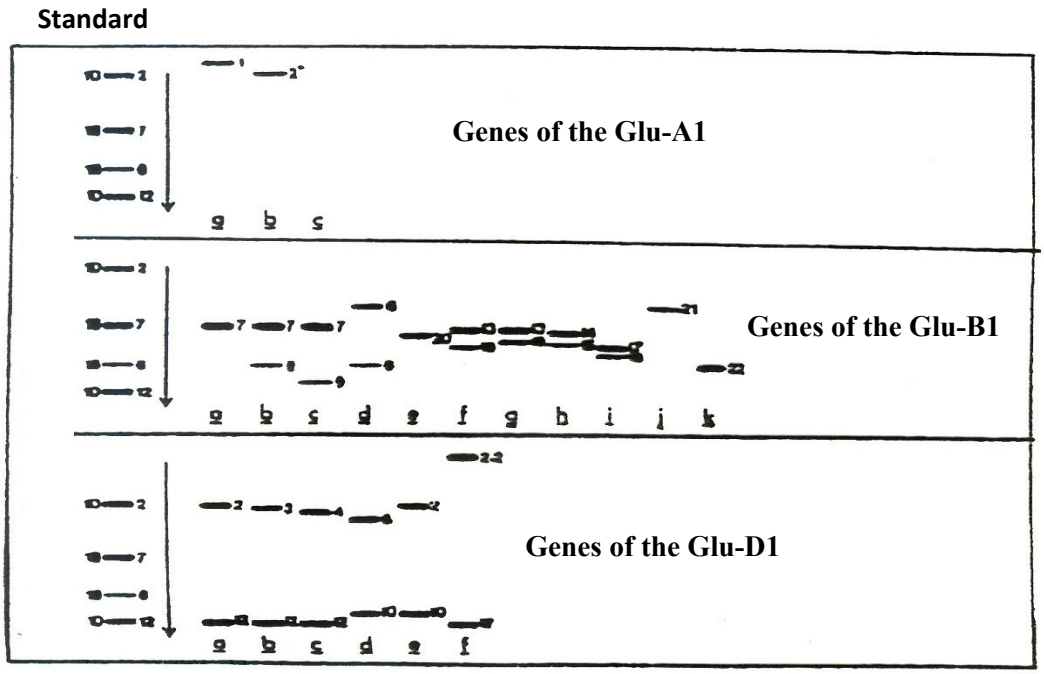


Figure 3.2 Basic catalog of alleles encoding HMW glutenin subunits (Payne et al., 1987)

The HMW glutenin subunits can be further classified into x-type and y-type based on their different molecular weight. Lower molecular weight is shown by the y-types in contrast with the x-type subunits, while the genes for the synthesis of the x-type proteins are located on Chromosome 1A, 1B and 1D and the genes for the synthesis of the y-types can be found on Chromosome 1B and 1D. It is known that the x-types are characterized by a low content of Cys residues (around 0.4 mol.%), while y-types show a high content of Cys residues (1.3 mol.%). The differences in the size of the x- and y-type subunits arise due to the different number of repeating nona- and hexapeptides in the y-type and nona-, hexa- and tripeptides in the x-type within the central domain. It is the repetitive central domain of the structure of HMW glutenin subunits that is represented by the x-type.

The HMW glutenin subunits are significant in terms of their relationship to the baking quality of flour. It has been proven that individual subunits, or their pairs, have different effects on technological quality. The subunits encoded by the Glu-D1 locus are the best studied. The pair of subunits 2+12 negatively affects bread-baking quality, and on the contrary, the presence of subunits 5+10 has a positive contribution to this quality. It has also been proven that the presence of the HMW glutenin subunits encoded by the Glu-A1 locus, i.e. subunits 2* and 1, increases breadmaking quality, and on the contrary, their absence, i.e. the presence of a null allele (an allele that does not encode an HMW glutenin subunit), decreases this quality.

There are certain alleles at the Glu-B1 locus whose products increase the breadmaking quality – alleles (17+18, 7+8, 7+9), and those (6+8 and 7) that fare worse in relation to the technological quality of wheat. The biochemical and physico-chemical properties of the HMW glutenin subunits and their relative representation have a more significant influence on flour quality than the total protein content. It has been proven that it is mostly the Glu-D1 locus that contributes to the variability in the technological quality of individual wheat varieties, while the effects of allelic variability on the Glu-A1 and Glu-B1 loci are manifested only in combination with the "high-quality" Glu-D1 allele, which encodes the subunits 5+10. The mechanism of these mutual interactions has not yet been elucidated. It is believed that a subunit with a greater effect on quality may have its cysteine residues in the protein molecule exposed

to the surface of the polypeptide, creating a disulfide bond between the subunits, which results in a strong and elastic gluten.

It is not the amount of HMW glutenin subunits but rather their representation that affects the prediction of technological quality of flour. Several criteria for a point evaluation of the individual loci were developed. The positive or negative contribution of the HMW glutenin subunits to bread baking can be expressed as a point value using the so-called Glu-score, Glu-ratings (Table 3.1).

Table 3.1 Contribution of some HMW glutenin subunits to the Glu-score value (Payne et al., 1987)

Glu-score value	**LOCUS**		
	Glu-A1	**Glu-B1**	**Glu-D1**
4	-	-	5+10
3	1	17+18	-
3	2*	7+8	-
2	-	7+9	2+12
2	-	-	3+12
1	0	7	4+12
1	-	6+8	-

The proven relationship between the Glu-score and technological quality has been established in selected varieties of wheat grown in Australia, Great Britain, Canada, Germany, Italy, the Netherlands, the Czech Republic and Slovakia (Figure 3.3). It has been proven that a lack of sulfur during grain formation negatively affects the quality of flour made from this grain, which is related to a change in the ratio between the individual protein fractions, and thus the content of the HMW glutenin subunits. The application of a nitrogen fertilizer generally increases the protein content in the grain, which also affects the technological quality of flour. A densitometric analysis of the electrophorograms of endosperm storage proteins shows that the amount of HMW glutenin subunits in relation to other storage proteins does not depend on the amount of fertilizer. The differences in protein content observed in the individual years of cultivation and at different locations in which the varieties were grown were the result of miniscule differences in the amount of individual types of glutenin subunits and in the amount of individual types of gliadins. The interactions between the genotype and the environment in determining the amount of HMW glutenin subunits were not identified. The ratio between the individual HMW glutenin subunits can be used to detect the genetic variability in the gene expression of individual HMW glutenin subunits, as this ratio does not depend on the location of the wheat crop.

DOI: https://doi.org/10.15414/2023.9788055226897

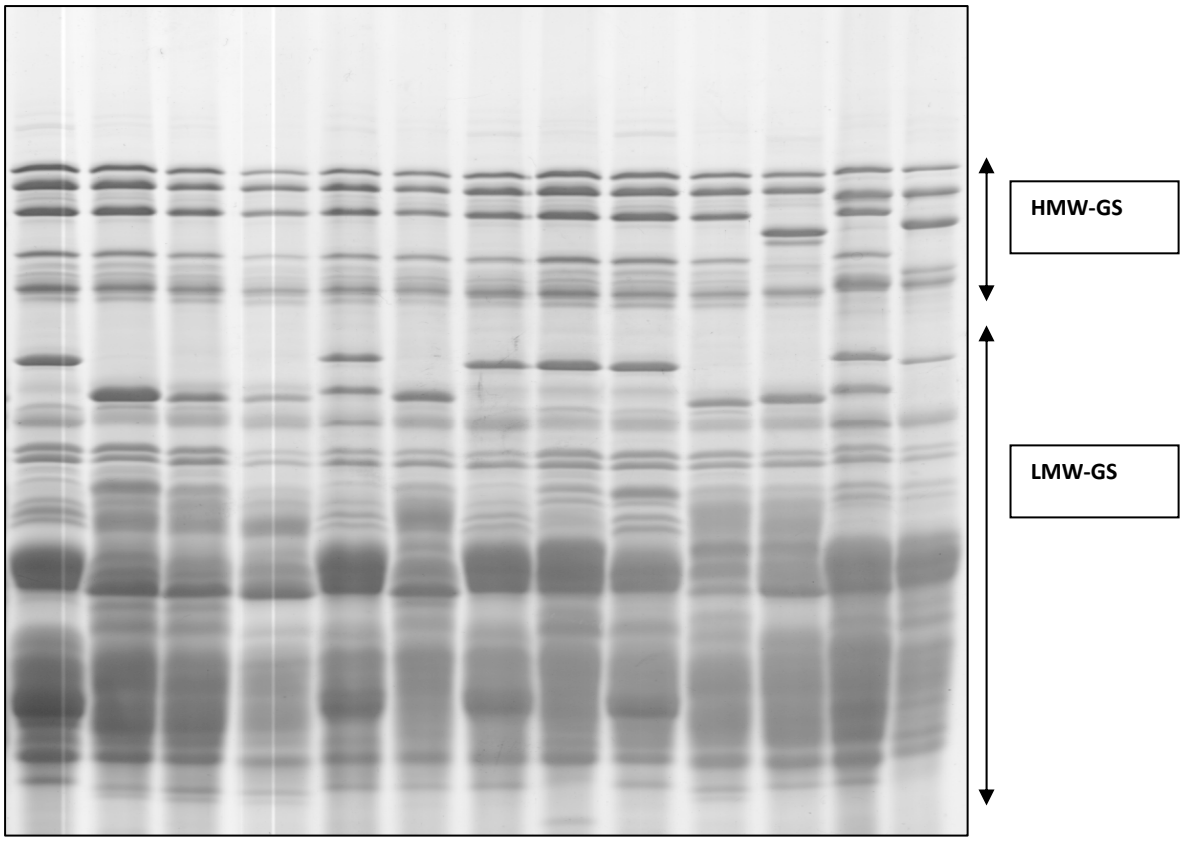


Figure 3.3 Electrophoretic separation of storage proteins of wheat grain (HMW-GS – high molecular weight glutenin subunits, LMW-GS – low molecular weight glutenin subunits)

The analysis of wheat storage proteins in A-PAGE can be used to determine the Gli 1B3 block, the presence of which (Figure 3.4) indicates that the wheat genotypes are characterized by low technological quality but an increased resistance to grass rust.

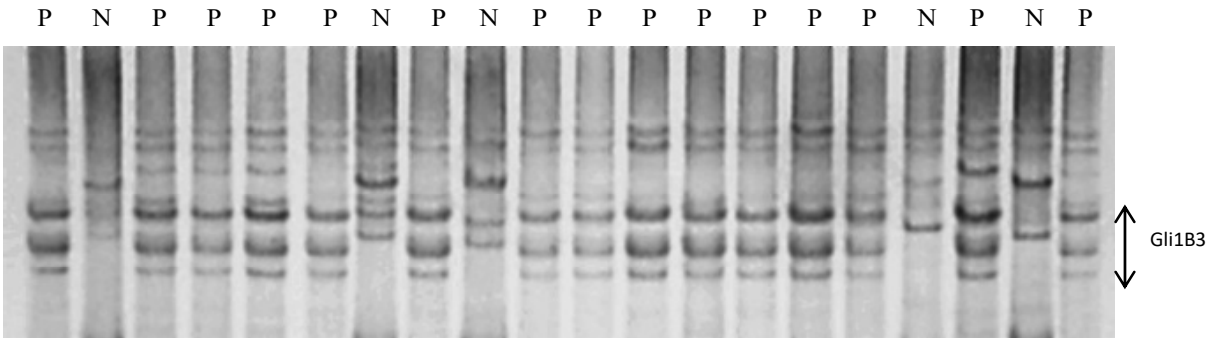


Figure 3.4 Electrophoretic spectrum of different wheat genotypes in polyacrylamide gel (A-PAGE). The Gli 1B3 block is marked with an arrow. Genotypes that contain the mentioned block are marked P.

DOI: https://doi.org/10.15414/2023.9788055226897

3.1.2 Genetic markers of barley

Storage proteins of barley (*Hordeum vulgare* L.) – hordeins (prolamins) are used as genetic markers for the identification of barley genotypes. They are suitable for distinguishing and identifying the individual barley genotypes (line, varieties), as well as for marking some economically important properties. Hordeins are very similar to gliadins in the percentage representation of the individual amino acids, and they contain a higher proportion of glycine, proline and valine, but less aspartic acid and/or asparagine.

a) Vertical electrophoresis in a polyacrylamide gel in the presence of sodium dodecyl sulfate made it possible to divide the hordein fraction into three groups: A with a molecular weight of 14,000 – 22,000 Daltons, B with a molecular weight of 28,000 – 49,000 Daltons and C with a molecular weight of 50,000 – 80,000 Daltons. The use of 2-mercaptoethanol to reduce disulfide bonds made it possible to detect a fourth group with a molecular weight greater than 100,000 Daltons, marked with the letter D, with the use of the above-mentioned electrophoretic method.

The Group A proteins are significantly different from typical prolamins in terms of their physico-chemical properties (molecular weight, isoelectric points, solubility) and amino acid composition (higher content of lysine, significantly lower content of glutamine and proline). Also, the genes controlling their synthesis are not located on Chromosome 5, but on Chromosomes 1 and 4.

The Group D proteins are composed of polypeptides with a relatively high molecular weight and higher representation of glycine residues, which are more similar to the subunits of wheat glutenins with a high molecular weight. The genes controlling their synthesis are located in the longer arm of Chromosome 5.

It follows from the above that only hordein Groups B and C can be considered typical hordeins, and both groups show a high degree of polymorphism. They differ from each other not only in molecular weights, but also in the presence of cystine – hordein B contains approx. 2.5% and hordein C only trace amounts.

b) Vertical electrophoresis in a starch gel made it possible to achieve a significant division of hordeins into three fractions: A, B and F hordeins. Minor components of hordeins C and D are found between zones A and B hordeins.

The loci determining hordeins A, B, F, C, D and E are localized on the short arm of Chromosome 5 in the following order: Hrd A, Hrd B, Hrd F, Hrd C, Hrd D, Hrd E. There is also a hordein locus labelled Hrd G, which determines the occurrence of two polypeptides in the B region of hordein. The Hrd G gene is strongly linked to the Hrd A locus ($p = 0.70$ to 2.56 Morgans).

The hordein components of the zones in the A hordein electrophoretic spectrum are inherited in blocks. At the same time, the existence of numerous allelisms of the Hrd A locus was confirmed. The individual blocks of the B hordein components are genetically determined by the allelic variants of the complex Hrd B locus. Likewise, F hordeins are inherited similarly to A and B hordeins.

The hordein loci Hrd C, Hrd D, Hrd E and Hrd G condition the occurrence or absence of individual minor, low-intensity components in the zone of low mobility of the electrophoretic spectrum of hordeins. These loci do not have such a complex, polycistronic structure as the Hrd A and Hrd B and/or Hrd F loci. They probably consist of only one gene in the dominant or recessive (null zone) state.

The Hrd A, Hrd B and Hrd F loci encode multiple types of proteins. The largest number of intense components was found in the area of the hordein spectrum corresponding to the Hrd F locus. If each hordein component of the hordein electrophoretic spectrum is genetically

determined by at least one gene, a cistron, the hordein locus controlling the synthesis of the block of hordein components, must be composed of several cistrons.

The designation of hordein component blocks contains the symbol HRD, which corresponds to the symbol of the hordein locus Hrd. Next, the designation of the hordein block contains a letter specifying the hordein locus or hordein type (A, B, F, etc.). Finally, the designation of the block of hordein components indicates the number of the allelic variant of the block of hordein components.

Cultivated barley (*Hordeum vulgare* L.) has only seven chromosomes in the haploid chromosome set of gametes. It is a typical diploid, probably originating from Hordeum spontaneum. As a result of the connected ears, the localization of hordein genes in Chromosome 5 of barley with other loci, genes and blocks of hordein components can act as genetic markers of other traits and properties bound to the mentioned binding group. The use of hordeins as genetic markers is made possible by the high heritability of the composition of electrophoretic hordein spectra.

Significant correlations between the variability of allelic block variants of hordein components and the variability of other traits and characteristics:

- Marking the resistance of barley to pathogens
- Marking the resistance of barley against frost
- Marking of barley grain production
- Marking of protein structure
- Marking of malting quality

The genes designated as Hrd-A3, Hrd-B4 and Hrd-F3 mark higher frost resistance of winter barley, while the Hrd-B1 gene is related to higher grain production and higher thousand-grain weight.

The allelic variants of hordein blocks can also mark the composition of barley grain proteins. The biotypes of "Oksamit" winter barley with the Hrd A1, Hrd B1, Hrd F1 and Hrd A3 genes are characterized by the same high protein content of the grain, a lower proportion of hordeins and a higher representation of glutelins, which was manifested in a higher content of lysine.

There is a correlation between the allelic variability of hordein loci and the malting quality of barley grain. The genotypes with Hrd A2 B21 or Hrd A2 B19 blocks were characterized by higher malting quality than the genotypes with the Hrd A2 B17 or Hrd A2 B8 hordein blocks. The polymorphism and specificity of hordeins are a prerequisite not only for the marking of individual genes found in the linkage with the hordein loci in chromosome 5, but also for the marking of uniform genotypes of barley lines and varieties.

Various electrophoresis methods were used to verify the barley varieties according to the electrophoretic composition of hordeins (electrophoresis of hordeins in starch gel, polyacrylamide gel, polyacrylamide gel in the presence of sodium dodecyl sulfate, isoelectric focusing, isoelectric focusing in ultrathin layers, etc.).

Compared to the identification of wheat varieties using gliadin electrophoresis, the lower discrimination ability of hordein electrophoresis in the identification of barley varieties is mainly affected by the high degree of genetic relatedness of current intensive barley varieties and the localization of hordein genes in one chromosome. Likewise, the overall variability of hordeins, expressed by the number of hordein components, is lower than the variability of gliadins.

The simultaneous use of hordein electrophoresis methods and some isozyme systems is therefore advantageous for the identification of barley varieties.

Hordein electrophoresis can also be used to study intravarietal hordein polymorphism. Some varieties of barley are heterogeneous in the composition of hordeins, consisting of two or more hordein lines. The knowledge of intravarietal hordein polymorphism refines the use of hordein-heterogeneous barley varieties as parental forms in the hybridization programs and makes it possible to rationally organize the maintenance breeding of hordein-heterogeneous barley varieties.

DOI: https://doi.org/10.15414/2023.9788055226897

4 DNA AS A MOLECULAR MARKER

The molecular markers revealing polymorphism in a DNA sequence have several advantages over the protein markers (storage proteins, isozymes). The DNA analyses help to monitor polymorphism in both coding (exons) and non-coding (introns) DNA sequences, while protein analyses only depend on the expression of coding sequences of the genome. Polymorphism can also be observed in pseudogenes and retrotransposons (mobile genetic elements), which are also interesting sequences in DNA.

The DNA markers are not affected by the stage of plant development or agro-ecological growing conditions. In contrast, some isoenzyme markers are dependent on the developmental stage of the plant, they must be analysed only in specific tissues, and their expression can also be conditioned by the agroecological conditions of cultivation. With the help of DNA markers, the genotype of plants can be determined at a very early stage of development with a minimal but sufficient amount of DNA. The determination of the nucleotide order of defined homologous sections of DNA from different genotypes and its comparison is the most direct monitoring of polymorphism at the DNA sequence level.

DNA polymorphism can be divided into:

1. point polymorphism, which is most often caused by a mutation of the corresponding nucleotide in the DNA sequence – nucleotide substitution or deletion, which can be observed during digestion with restriction endonucleases, which is used in the RFLP technique to study the length polymorphism of restriction fragments,

2. polymorphism in the number of tandem repeats (VNTR – Variable Number of Tandem Repeats) divided into several groups based on the length of the repetitive motif and the number of copies.

Tandem repeats can be divided as follows:

Satellites – highly repeated segments composed of 100 or more (up to 300) repeated nucleotides, which are more or less uniform and have a length of 103 to 107 nucleotides.

Minisatellites – moderately repeated segments of 10 to 100 nucleotides forming more or less uniform clusters 102 to 105 nucleotides long.

Microsatellites – short stretches of 2 to 10 repeating nucleotides grouped by up to 102 nucleotides.

Mononucleotides – uniform single nucleotide sequences of any length.

The microsatellites or short tandem repeats (STR-Short Tandem Repeats), simple sequence repeats (SSR-Simple Sequence Repeat), simple sequence length polymorphism (SSLP-Simple Sequence Length Polymorphism) are divided into **perfect**, **imperfect** and **compound microsatellites**. Perfect microsatellites are formed by one continuous repetitive motif, e.g. $(AT)_{14}$. In imperfect microsatellites, the basic motif is interrupted by a sequence of

several nucleotides, e.g. $(AC)_{16}GT(AC)_4$. Composite microsatellites represent several sequence motifs, e.g. $(AT)_{15}(AC)_{20}$. As a result of new information about the different sequence motifs within the microsatellites, Chambers and MacAvoy (2000) expanded the above division into six classes (Table 4.1), replacing the term "perfect" with the term "pure" for perfect microsatellites and replacing the term "imperfect" with "interrupted pure" to denote imperfect microsatellites.

Table 4.1 Six classes of microsatellites (Chambers and MacAvoy, 2000)

Class	**Sequence**
Pure	-(AT)16-
Interrupted pure	-(TA)-(CA)4-TA-(CA)6-
Compound	-(AT)12-(AC)8-
Interrupted compound	-(AC)14-AG-AA-(AG)12-
Complex	(TTTC)3-4-(T)6-(CT)0-1- (CYKY)n-CTCC-(TTCC)2-4
Interrupted complex	An allele with a break inside a repeat unit

Microsatellites occur frequently, randomly and in high abundance in all eukaryotic nuclear DNA examined. The frequency of occurrence of microsatellites varies significantly across the organisms. It is estimated that the human body contains on average 10 times more microsatellites than the plant genome. Dinucleotide repeats (AC)n and (GA)n most often occur in plant species and e.g. wheat contains more dinucleotide repeats than rice or corn.

Trinucleotide and tetranucleotide repeats have also been found in the plant genome. Of these, (AAG)n and (AAT)n are the most common.

Most of the repetitive DNA sequences are found in the non-coding region of DNA, i.e. in introns. It was found that the exon region contains more trinucleotide repeats than dinucleotide repeats. Repetitive sequences of wheat and other Poaceae genomes (barley, rye) make up to 80% of the genome. The length and number of repetitive units varies.

EST sequences

To create expressed sequences (Expressed Sequenced Tag - EST), it is necessary to construct a library containing short sections of DNA coding obtained by isolating mRNA and the subsequent cDNA synthesis. By sequencing the end sequences of cDNA fragments, information is obtained for the creation of EST probes. Inside the EST sequences, there are many trinucleotide repeats, but also repeats with higher and lower repetitive motifs. An international database of EST sequences of the Triticeae (ITEC) family has been created, which contains the EST sequences of wheat.

Current DNA polymorphism analyses are based either on DNA hybridization (**hybridization techniques**) or the principle of DNA amplification by PCR (**amplification techniques**) (Table 4.2).

The DNA techniques based on the principle of hybridization (RFLP, etc.), i.e. hybridization techniques, were termed the DNA techniques of the **first generation**. The

techniques based on the principle of PCR reaction (RAPD, SSR, AFLP, etc.) and all amplification techniques listed in Table 3.2 are referred to as the DNA techniques of the **second generation**. Scientists include the SNP and DArT in the DNA techniques of the **third generation**. An overview of the DNA techniques is provided in Table 4.2.

Table 4.2 Overview of DNA polymorphism detection techniques (Kraic, 1999, modified).

DNA polymorphism detection techniques	
1. hybridization	**RFLP** (Restriction Fragment Lenght Polymorphism) **DNA fingerprinting**
2. amplification **a) random primers**	**RAPD** (Randomly Amplified Polymorphic DNA) **DAF** (DNA Amplification Fingerprinting) **AP-PCR** (Arbitrarily Primed – PCR)
b) semi-random primers	**AFLP** (Amplified Fragment Lenght Polymorphism) **MP-PCR** (Microsatellite Primed – PCR) **ISSR** (Inter-SSR Amplification)
c) specific primers	**STS** (Sequence Tagged Sites) **SCAR** (Sequence Characterized Amplified Region) **CAPS** (Cleaved Amplified Polymorphic Sequences) **IRAP** (Inter-Retrotransposon Amplified Polymorphism**)** **EST-SSR** (Expressed Sequence Tag-SSR) **STMS** (Sequence Tagged - Microsatellite Sites) **SCoT** (Start Codon Targeted)
d) combination of different primers	**RAMP** (Random Amplified Polymorphic Microsatellites) **SAMPL** (Selective Amplification of Microsatellite Polymorphic Loci) **REMAP** (Retrotransposon-Microsatellite Amplified Polymorphism)
3. amplification – hybridization	**RAMPO** (Random Amplified Microsatellite Polymorphism)

4.1 Hybridization techniques for searching for DNA polymorphism

The **DNA polymorphism hybridization techniques** reveal the differences between the genotypes or species in different lengths of DNA fragments.

1. C**leaved by restriction endonucleases** (RFLP technique), with the changes arising as a result of a point mutation

2. Due to the difference in the number of tandem repetitions, the so-called **VNTR** (Variable Number of Tandem Repeats) loci (oligonucleotide fingerprinting)

4.1.1 RFLP technique

The **RFLP technique** (Restriction fragment length polymorphism) was developed as the first technique to detect polymorphism at the DNA level. In the beginning, it was used primarily to map the human genome, but later it was also used in the mapping of plant and animal genomes.

The RFLP analyses include the following steps (Figure 4.1):

1. Isolation of DNA

2. Cleavage of DNA into fragments using restriction enzymes (endonucleases)

3. Separation of DNA fragments in gel electrophoresis (agarose) according to size

4. Transfer of DNA fragments to a nylon or nitrocellulose membrane filter (Southern blotting)

5. Hybridization of DNA fragments with labelled probes (radioactive or chemiluminescent)

6. Detection of hybridized molecules (autoradiography with radioactive labelling), evaluation of results

DOI: https://doi.org/10.15414/2023.9788055226897

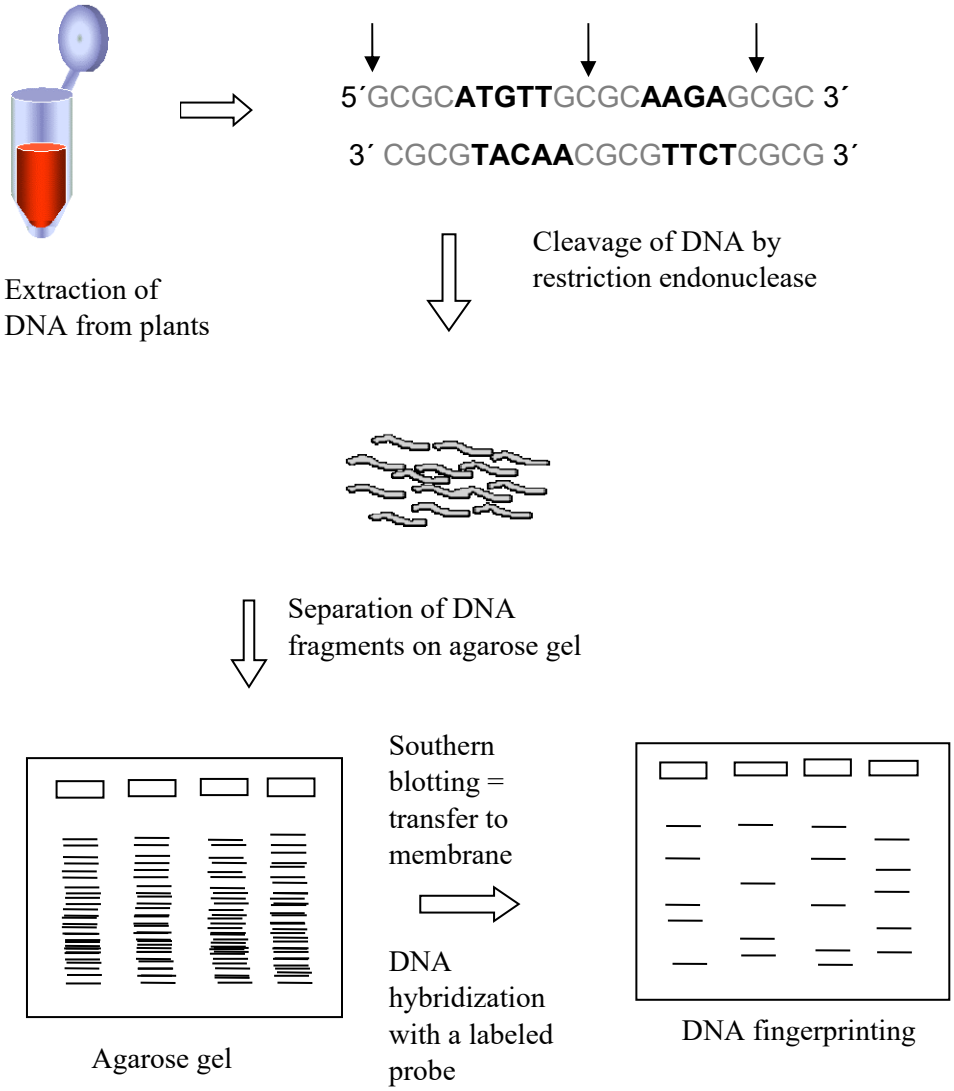


Figure 4.1 Principle of the RFLP technique.

The DNA restriction fragment length polymorphism (RFLP) is a consequence of the diversity that arises from the cleavage of restriction sites recognized by a specific restriction endonuclease. It can be a point mutation that results in the loss or gain of a restriction site, an insertion or deletion of DNA between two restriction sites, or a deletion spanning the restriction site. Restriction endonucleases recognize short (3-9 bp) sequences and cleave DNA at a precise site, usually palindromic. After the DNA cleavage by restriction endonucleases, a set of fragments of different lengths is formed. If one base pair changes at the cleavage site, the DNA is cleaved by the enzyme at another site.

After the cleavage of DNA by restriction endonucleases, separation of the cleaved DNA fragments electrophoretically in agarose gel is performed, followed by the transfer of DNA from the gel to a hybridization membrane (Southern blotting), hybridization with a labeled probe (radioactive or chemiluminescent) and detection of hybridized molecules.

The probes used in the RFLP analyzes can be locus-specific or multicopy. To differentiate the genetically related species, it is necessary to select the probes that have

DOI: https://doi.org/10.15414/2023.9788055226897

a sufficiently large polymorphism. Such probes are searched in a genomic library or in a cDNA library. By using a higher number of different probes and restriction endonucleases, it is possible to increase the efficiency of the RFLP technique. The RFLP markers are codominantly inherited.

Advantages and use of the RFLP technique:

• Construction of RFLP genetic maps – rice, tomato, barley, corn, wheat and others

• Identification of genotypes and varieties of plants in any tissue at any stage of plant development, regardless of the influence of external agroecological conditions

• Codominance of RFLP markers allows us to distinguish a homozygote from a heterozygote

• Differentiation of species or populations (single-locus probe) or individuals (multilocus probes)

• Knowledge of the DNA sequence of the studied genome is not required

• Indirect selection of genes using qualitative features – tightly linked RFLP markers that can be converted to PCR markers, such as: STS (Sequence tagged site) and CAPS (Cleaved amplified polymorphic site)

Disadvantages of the RFLP technique:

• Requires a large amount of high-quality pure DNA for digestion with restriction enzymes

• A relatively large number of probes are required for a reliable detection of diversity between the genotypes, which is time-consuming and financially demanding

• The analyses are time-consuming and financially demanding (with radioactive marking)

• A lower degree of polymorphism was observed compared to other markers (microsatellites, AFLP), e.g. due to the extensive genome size in wheat, polyploid nature of the wheat genome and a high proportion of repetitive DNA sequences

4.1.2 DNA fingerprinting

DNA fingerprinting (Oligonucleotide fingerprinting, DNA typing, DNA profiling) is similar to the RFLP technique and it is based on the principle of DNA hybridization. After the digestion of DNA with specific restriction endonucleases, the cleaved fragments hybridize with synthetic oligonucleotide probes, which are complementary to the microsatellites or simple sequence repeats (SSRs). Polymorphism obtained by oligonucleotide fingerprinting is formed on the basis of different lengths of restriction fragments that contain microsatellites.

The cleaved fragments that hybridize to the synthetic oligonucleotides vary considerably from probe to probe. The hybridized fragments range in size from a few hundred base pairs up to 8-10 kb. Since simple sequence repeats (SSRs) are frequently found in DNA

and are uniformly distributed throughout the genome, a large number of hybridized fragments varying in length can be obtained.

When detecting DNA polymorphism in several loci at the same time, multilocus probes are used, i.e. even a small set of different probes is sufficient to cover a certain characteristic part of the genome.

Advantages and uses of oligonucleotide fingerprinting:

• multilocus probes can be used to identify and characterize the varieties, species and breeding lines

• Analysis of links between the genes

• Determination of genetic relatedness between the genotypes

• High polymorphism obtained by this technique was observed between the related genotypes

• Use in various areas of genomic analysis, such as paternity testing, genotype identification and population genetics

Disadvantages of oligonucleotide fingerprinting:

• Lower polymorphism was observed compared to other techniques

• Impossibility to detect small fragments containing microsatellites

4.2 Amplification techniques to search for DNA polymorphism – PCR markers

The molecular markers based on the PCR principle have several advantages over the already mentioned hybridization techniques (RFLP, DNA fingerprinting). These primarily include high automation and shorter analysis time, only a small amount of DNA, extreme sensitivity of the PCR reaction and, last but not least, lower financial costs. **PCR,** or **Polymerase Chain Reaction,** was first interpreted by Saiki in 1985, which also marked the beginning of the development of amplification techniques (Figure 4.2).

PCR is a method in which a selected section of DNA is multiplied several times by the action of a specific thermostable DNA polymerase and other components of the reaction mixture. The analysis consists of three cyclically repeated steps that include denaturation, annealing and polymerization. Primers are used in the PCR reaction. These are either randomly synthesized short sections of DNA or specific sequences of nucleotides to known DNA sequences, which, after DNA denaturation, are attached to complementary DNA sites, which starts the synthesis of the second strand of the selected DNA section. The whole process takes place in a thermocycler, which can change the temperature level in a very short time. The process takes place in several cycles (25-40), with each cycle doubling the amount of DNA. The resulting number of copies is equal to 2n, where n is the number of cycles.

The amount of DNA copies obtained in this way is detected after electrophoresis in an agarose gel under a UV lamp, or by autoradiography, silver staining or fluorescence after electrophoresis in polyacrylamide (PAGE) gels.

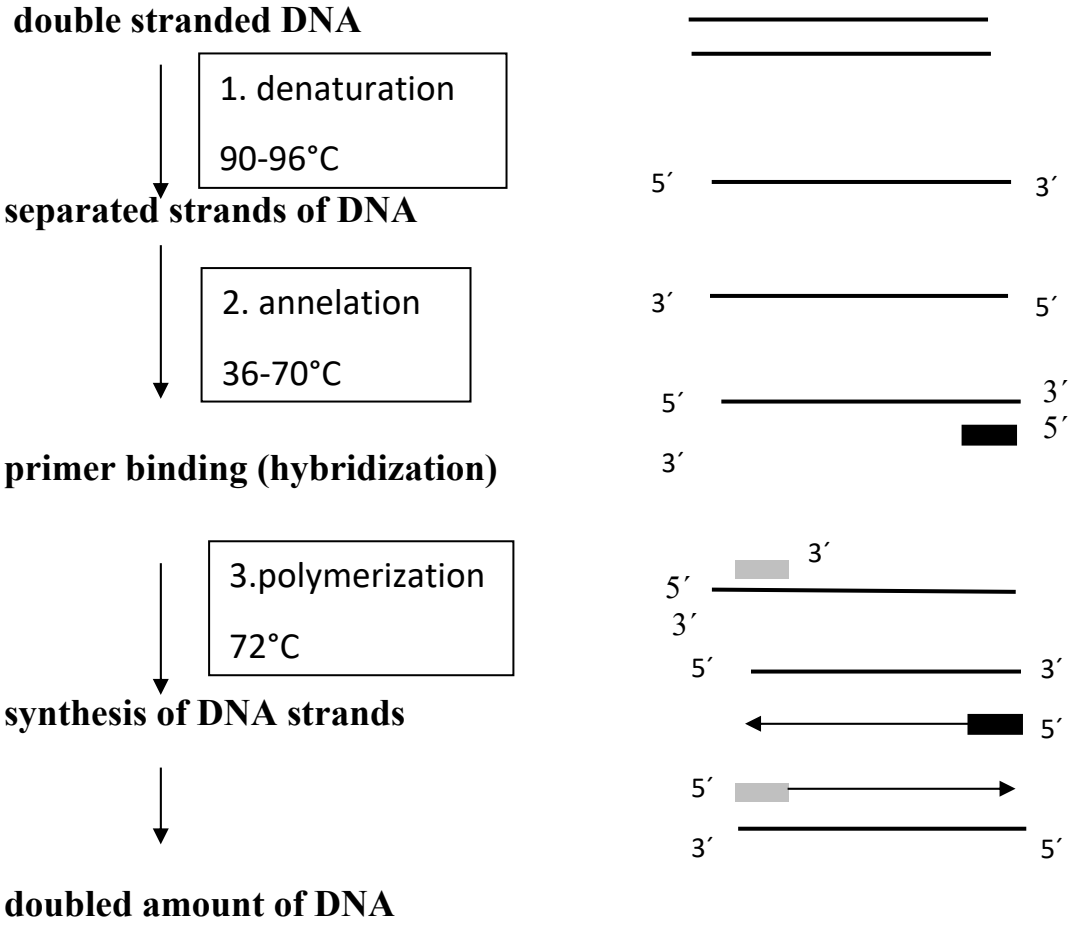


Figure 4.2 Polymerase chain reaction (PCR).

Each amplification technique has its advantages and disadvantages, its applicability often depends on the size of the genome, but time constraints and financial demands are also a major factor at play when choosing a particular technique. An overview of selected amplification techniques with their advantages and disadvantages is presented in Table 4.3 and they are compared with the hybridization technique (RFLP).

DOI: https://doi.org/10.15414/2023.9788055226897

Table 4.3 DNA techniques and their characteristics (Semagn et al., 2006, modified)

Factors	RFLP	RAPD	STMS	AFLP	SCoT	SNP	EST
amount of DNA	big	small	small	small	small	small	small
PCR method	no	yes	yes	yes	yes	yes	yes
radioactive marking	yes, no	no	yes, no	yes, no	no	yes, no	yes, no
locus specificity	yes	no	yes	no	no	yes	yes
required sequence information	no	no	yes	no	no	yes	yes
dominance (D)/ codominance (K)	K	D	K	D,K	D	K	K
reproducibility	high	difficult	high	high	high	high	high
level of polymorphism	average	average	high	average	average	average-low	average
another advantage	large number of accessible RFLP probes	cheap, technically undemanding	locus specific, possibility of automation	detection of a large number of loci in one reaction	cheap, technically simple, mapping of coding sequences	detection of SNPs in specific genes using EST databases	coding sequence mapping

Amplification techniques use:

1. Random primers

• **RAPD (Randomly Amplified Polymorphic DNA)** – one random primer usually 10 nucleotides long

• **DAF (DNA Amplification fingerprinting)** – one random primer approximately 5-8 nucleotides long

• **AP-PCR (Arbitrary Primed-Polymerase Chain Reaction)** – one primer with a length of 10-50 nucleotides

2. Semi-random primers

• **ISSR (Inter Simple Sequence Repeat)** – primers contain random repetitive sequences with the addition of other (anchoring) sequences at the 3' or 5' end of the primer

• **AFLP (Amplified Fragment Length Polymorphism)** – adapters are attached to the restriction fragments, to which AFLP primers are complementary

• **MP-PCR (Microsatellite-Primed PCR)** – primers contain randomly selected repetitive sequences,

• **RAMP (Random Amplified Polymorphic Microsatellites)** – a combination of a microsatellite primer anchored at the 5' end and a random RAPD primer with a length of 10 nucleotides

• **SAMPL (Selective Amplification of Microsatellite Polymorphic Loci)** – a combination of two primers is used, while one of the primers is self-anchored at the 5' end of a mixture of microsatellite repeats and the other primer is an AFLP primer, which is labeled based on the sequences of the synthetic adapter and the restriction site, and carries two to three selective nucleotides

• **REMAP (Retrotransposon-Microsatellite Amplified Polymorphism)** – a combination of two primers is used, one of which is complementary to the repetitive sequences anchored at the 3' end and the other is specific to the LTR sequences of retrotransposons

• **RAMPO (Random Amplified Microsatellite Polymorhism)** – random RAPD primers are used for amplification and then the PCR products are hybridized with a radioactively labeled microsatellite probe

3. Specific primers derived from known DNA sequences

• **STS (Sequence Tagged Site)** – primers are obtained by sequencing fragments that hybridize with the RFLP probes and are bound to the desired character.

• **STMS (Sequence Tagged Microsatellite Site)** – primers are obtained by DNA sequencing in places that border on the microsatellites in specific loci

• **SCAR (Sequence Characterized Amplified Region)** – primers are obtained by sequencing the end sequences of RAPD or ISSR fragments,

• **CAPS (Cleaved Amplified Polymorphic Sequences)** – primers are synthesized on the basis of cDNA sequences or cloned RAPD fragments, but the resulting PCR products are further cleaved by restriction endonucleases,

• **IRAP (Inter-Retrotransposon Amplified Polymorphism)** – primers are synthesized to complementary LTR (Long Terminal Repeats) sequences of retrotransposones,

• **EST-SSR (Expressed Sequence Tag – SSR)** – primers are obtained by the DNA sequencing of the bordering microsatellites, which are located in the coding DNA sequences,

• **SCoT technique (Start Codon Targeted)** – primers are designed in short conserved regions and contain the ATG start (initiation) codon in the plant genes.

4.2.1 Visualization of amplified DNA fragments

Agarose gels
• Detection of DNA fragments ranging in size from tens to thousands of base pairs (bp) with a resolution greater than 3 bp
• Visualization of DNA using ethidium bromide (EtBr) as a fluorescent intercalating agent, which is incorporated between the DNA strands and moves with them
• EtBr fluorescence indicates the positions of DNA molecules under ultraviolet light

Polyacrylamide gels
• Identification and separation of small molecules and DNA fragments (50-500 bp) with a resolution of 1 bp
• Separation takes place in electrode buffer solutions exposed to high voltage (2,000-3,000 V) or higher constant power (50-70 W) between the electrodes
• Separated DNA molecules are visualized using vertical electrophoresis in denatured and non-denatured PAGE gels using silver nitrate, radioactively or fluorescently

4.2.2 RAPD technique

RAPD markers (Random amplified polymorphic DNA) were first described independently by Williams et al. (1990) and Welsh and McClelland (1990). These are simple, usually 10 bp primers that bind randomly to complementary sites throughout the genome (Figure 4.3). Amplification uses an annealing temperature of around 35 – 38°C. One RAPD primer allows the amplification of several fragments corresponding to several loci, which are mostly dominant. However, it is not possible to distinguish dominant homozygotes from heterozygotes using RAPD primers.

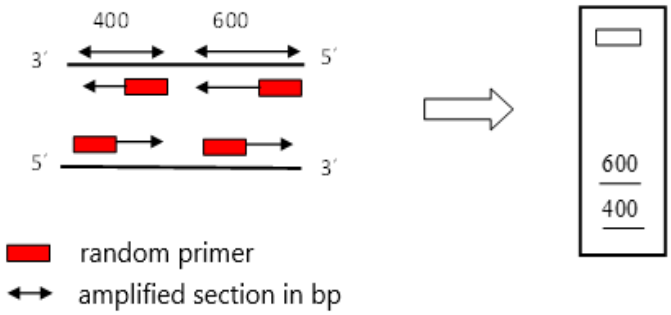


Figure 4.3 Principle of the RAPD technique

The largest inter-individual differences using the RAPD technique were observed in cases where a base pair substitution, insertion or deletion occurs at the primer binding site. It has been proven that the RAPD method can be successfully used in the study of plant taxonomy, determination of systematic relationships and identification of parents.

The RAPD markers have a use case in almost all plant species, especially in cereals, oilseeds and conifers. In wheat, the RAPD technique was used by Devos and Gale already in 1992. However, similarly to the RFLP markers, a low degree of DNA polymorphism and low reproducibility of results was detected in wheat, which was attributed to the large size of the wheat genome and to the high proportion of repetitive sequences in DNA, which represents up to 80% of the genome in wheat.

Advantages and use of the RAPD method:
• Simple method, does not require knowledge of the genome sequence
• Used in the construction of genetic maps (*Arabidopsis, Heliantus*...)
• Mapping of monogenic and polygenic traits – indirect selection of the fission population during plant breeding programs
• Determination of genetic relatedness between the genotypes, evaluation of plant genetic resources, fingerprinting of individuals

Disadvantages of the RAPD method:
• Low reproducibility
• Short primers cause changes in the electrophorogram (missing band) and are sensitive to changing conditions

DOI: https://doi.org/10.15414/2023.9788055226897

4.2.3 AFLP technique

The AFLP technique (Amplification fragment length polymorphism) was created by Vos et al. in 1995. It is a highly sensitive genome-wide method to detect polymorphism. The AFLP technique does not require knowledge of the genome sequence. It is a technique with high multiplex ratios, which is based on the principle of selective PCR amplification of restriction fragments cleaved by specific restriction endonucleases and includes the following steps (Figure 4.4):

1. DNA isolation
2. DNA is usually cleaved by two restriction endonucleases (*PstI/MseI, EcoRI/MseI* and others)
3. Specific double-stranded DNA adapters are attached to the DNA restriction fragments
4. Selective amplification by means of primers that are labeled to contain sequences complementary to both adapters and restriction sites – usually one to three selective nitrogenous bases are attached to the 3'-end of the primer, thereby amplifying only part of the restriction fragments, which, however, generate a sufficiently large number of fragments enabling the detection of DNA polymorphism
5. Amplified products are separated in denatured PAGE gels and visualized radioactively, fluorescently, or with silver staining

Silver staining has several disadvantages compared to radioactive or fluorescent marking of primers. The first disadvantage of silver staining is often the different intensity of the bands at different sizes of the DNA fragment, especially small fragments in the range of 80-100 base pairs. The second disadvantage is the presence of a duplex due to the staining of both strands of the same DNA fragment, which can move at different speeds, which is not a problem with alternative approaches since only one primer is labeled in them.

The AFLP technique belongs to the group of methods with a high multiplex ratio. 50-100 fragments are amplified per reaction, of which up to 50% can be potential markers. A point mutation or insertion or deletion of a nucleotide that occurs randomly throughout the genome serves as the basis for the detection of AFLP polymorphism.

In general, the AFLP markers that have been created by combining the different restriction enzymes are freely scattered throughout the genome and affect the degree of polymorphism. In wheat, a higher degree of polymorphism was detected using the *PstI/MseI* restriction enzyme combination than the *Eco*RI/*Mse*I enzyme combination. For the analyses of larger genomes, such as wheat, the presence of primers containing 3 selective nucleotides is required during the selective amplification, while for relatively smaller genomes, such as rice, two selective nucleotides are enough to collect a sufficient number of DNA fragments. To obtain a reasonable number of DNA fragments, it is advisable to design primers that do not contain large amounts of AT nucleotides since AT nucleotides are common in large plant genomes such as wheat.

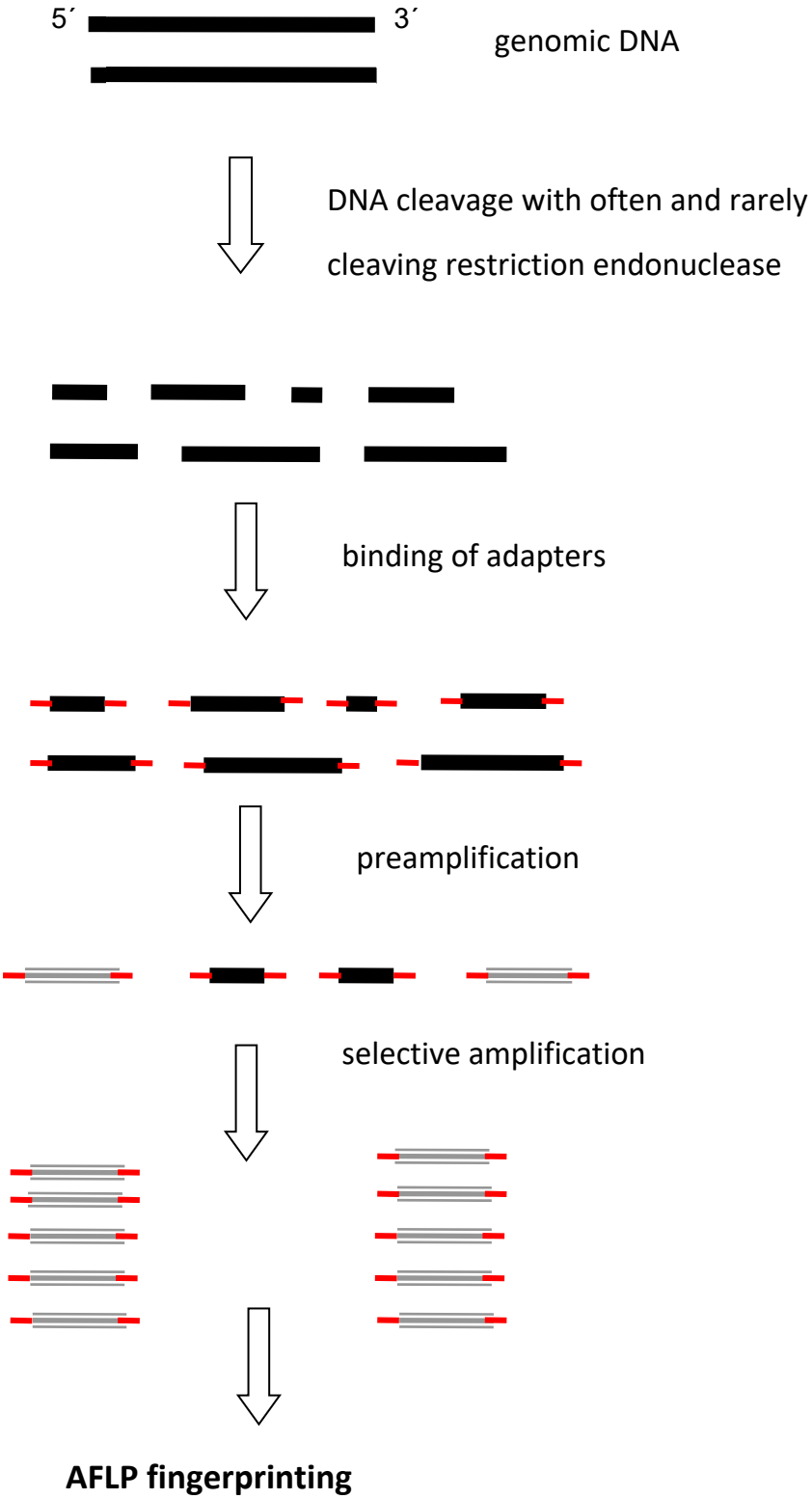


Figure 4.4 Principle of the AFLP technique

DOI: https://doi.org/10.15414/2023.9788055226897

Both dominant and codominant AFLP markers can be analyzed. The AFLP method also makes it possible to rapidly generate new markers either randomly scattered throughout the genome or linked to a specific region of the genome.

Advantages and use of the AFLP technique:

• Extremely sensitive
• Capable of distinguishing homozygotes from heterozygotes
• 50-100 fragments are amplified per reaction, of which up to 50% can be potential markers, it is highly sensitive and reproducible
• A point mutation, or the insertion or deletion of a nucleotide at the cleavage site serves as the basis for the detection of AFLP polymorphism
• AFLP technique is highly reproducible
• Suitable for detecting polymorphism throughout the genome, creating genetic maps, determining genetic relationships between the genotypes and identifying closely linked DNA markers
• AFLP markers can be converted to STS markers, which can be used in MAS thanks to the simplicity of their use compared to the AFLP technique

Disadvantages of the AFLP technique:

• Requires a larger amount of DNA (1 mg) than the RAPD method
• Technically and financially more demanding

4.2.4 STMS technique

The technique using microsatellites in which DNA polymorphism is detected by PCR in the individual loci using locus-specific primers flanking the microsatellite from both sides, was named **STMS** (Sequence tagged microsatellite sites) by Beckmann and Soller in 1990. The STMS markers are polymorphic in terms of the varying length of microsatellites at specific individual loci (Figure 4.5).

The first step in the creation of STMS markers requires cloning and sequencing and is therefore financially demanding and labor intensive. Once the locus-specific primers are developed, the method becomes efficient and very accessible. The differences in the length of amplified fragments are difficult to detect in the agarose gels stained with ethidium bromide, therefore differentiation without the use of radioisotopes can be carried out in polyacrylamide gels stained with silver, or fluorescently labeled primers can be used in combination with a semi-automatic DNA sequencer.

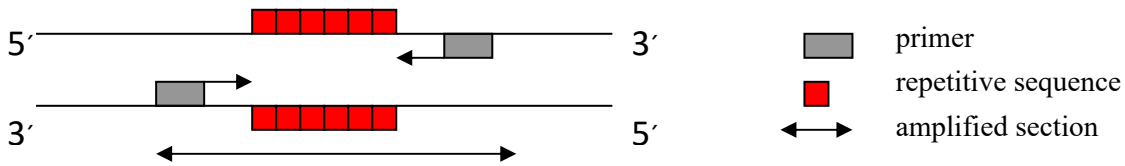


Figure 4.5 Principle of STMS technique

DOI: https://doi.org/10.15414/2023.9788055226897

Advantages and use of the STMS technique:
• Analyses are based on polymerase chain reaction (PCR)
• Highly polymorphic, highly sensitive and reproducible
• Suitable for genetic mapping and population studies due to the high frequency of microsatellites in eukaryotes
• Markers are genome-specific with codominant inheritance
• Useful tool for studying genetic relationships between the varieties, species and populations

Disadvantages of the STMS technique:
• Quite financially demanding
• Requires information about the DNA sequence for each marker in the individual loci, which is a disadvantage for genomes that are not well characterized yet
• Technically more demanding than e.g. the RAPD method

4.2.5 SCoT technique

The SCoT (Start Codon Targeted) technique was developed by Collard and Mackill in 2009 who proposed the SCoT technique for analysis. The SCoT technique uses short conserved regions in plant genes that surround the translation initiation start codon (ATG). The DNA markers are produced by the PCR reaction using one primer (18 nucleotides), which is designed to flank the ATG initiation codon on both sides, and which serves as both a forward and a reverse primer (Figure 4.6). Amplified fragments are separated by standard horizontal electrophoresis using agarose gels and dyes (ethidium bromide and others) with subsequent visualization under UV light. They are dominant marker systems for which genome sequence information is not required and polymorphism can be relative to gene functionality. The technique was initially verified on a model plant species – rice (*Oryza sativa*).

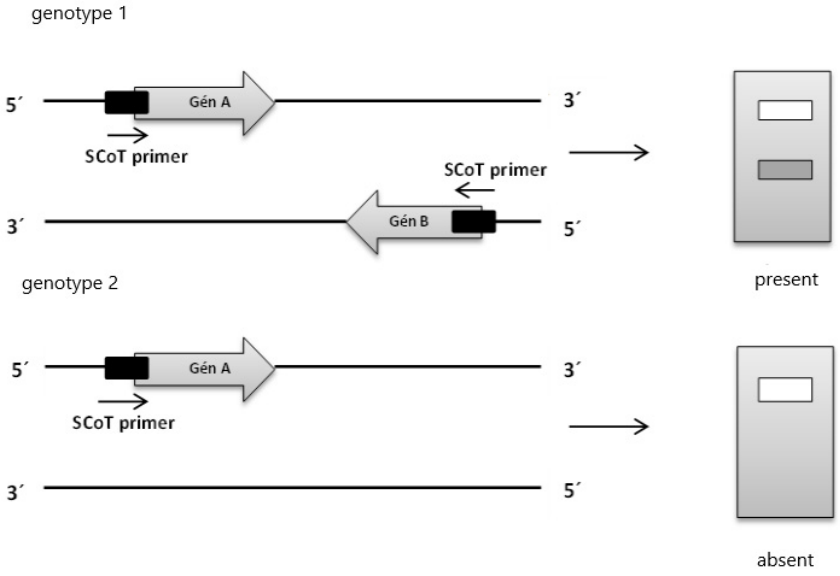


Figure 4.6 Principle of SCoT technique (Collard and Mackill, 2009, modified)

DOI: https://doi.org/10.15414/2023.9788055226897

Advantages and use of the SCoT technique:

- High reproducibility.
- Technical simplicity and financial simplicity.
- Use in QTL mapping.
- Possibility of incorporating the SCoT markers into existing genetic maps.
- Targeting specific chromosomal regions.

Disadvantages of the SCoT technique:

- Dominance of markers (it is not possible to distinguish a dominant homozygote from a heterozygote).

4.2.6 IRAP technique

Retrotransposons are a sizable class of repetitive DNA found in large quantities in plants, animals and fungi. They are numerous and the most active in plant genomes, constituting up to 40-60% of the total DNA.

Transposable (moving) elements can be divided into two large groups:
Type I mobile elements are called **retrotransposons.**
Type II mobile elements are called **DNA transposons.**

Type I retrotransposons are divided into 3 groups:
1) LINE (long-interspersed elements)
2) SINE (short-interspersed elements)
3) LTR (long terminal repeat)
They contain: a) copy-like elements
b) gypsy-like elements

The domains located near the LTR are conserved within certain groups and are therefore useful for designing primers in PCR reactions.

The IRAP technique (Inter-Retrotransposon Amplified Polymorphism) amplifies the regions between two nearby retrotransposons, or LTRs. The IRAP primers can be attached to both the 5' and 3' ends and they can point from the beginning of one LTR in one retrotransposon to the beginning of the LTR in another retrotransposon (head to head) (Figure 4.7) or from the end of one LTR in one retrotransposon to the end of the LTR in another retrotransposon (tail to tail).

The IRAP method uses a simpler horizontal agarose system and is one of the dominant marker systems.

Advantages and use of the IRAP technique:

- Highly polymorphic, highly sensitive, reproducible.
- Representation of retrotransposons in a large number of copies.
- Localization in different places of the chromosome.
- Suitable tool for genetic mapping and population studies due to the high frequency of retrotransposons in eukaryotes.
- Search for molecular markers (in connection with agronomically important properties).
- Useful tool for studying polymorphism and genetic relationships between the varieties, species and populations.

Disadvantages of the IRAP technique:

- Dominance of markers (it is not possible to distinguish dominant homozygote from heterozygote).

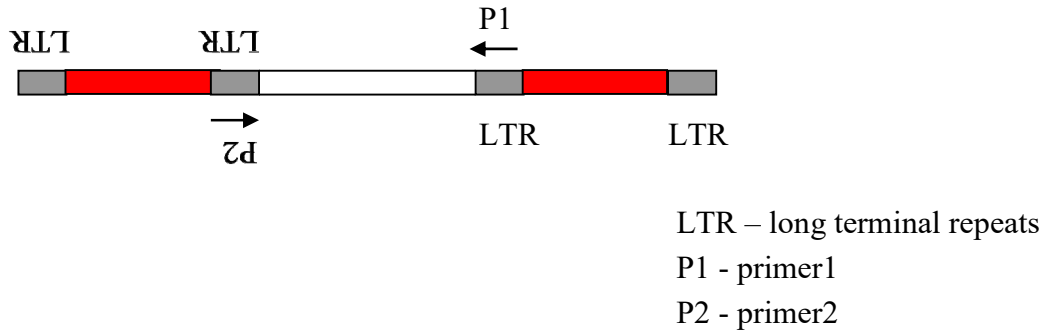


Figure 4.7 Principle of IRAP technique

4.2.7 SNP technique

The SNP (Single nucleotide polymorphism) technique is capable of capturing single nucleotide changes in DNA (single nucleotide insertions or deletions). With its help, it is also possible to distinguish related genotypes that other techniques cannot distinguish.

The identification of SNP markers depends on the information about the DNA sequence of the genomes. SNPs were discovered when working with a single gene during the comparison of the DNA sequence of different genotypes.

Currently, SNPs have been identified in large sequencing projects of individual plant genomes. SNPs are obtained by comparing the data from sequenced genomes and existing databases of EST sequences (cloned and sequenced EST sequences).

DOI: https://doi.org/10.15414/2023.9788055226897

4.3 Plant genomes mapping using molecular markers

Physical mapping – reveals not only the sequence of coding or non-coding DNA, i.e. the sequence and function of genes in a given genome, but also the mutual interactions of genes in the genome, or the gene-marker and marker-marker relations.
When mapping, it is possible to create 4 types of maps:

1. **Physical map** – physical arrangement between the genes, gene and marker, or markers (expressed in bp).
2. **Genetic map** – created based on the frequency of recombinations between the genes, gene and marker, or markers (expressed in cM) (Figure 4.8).
3. **Linkage map** – based on the genetic map, it indicates the genetic relationships.
4. **Cytological map** – position of genes or markers on the chromosome.

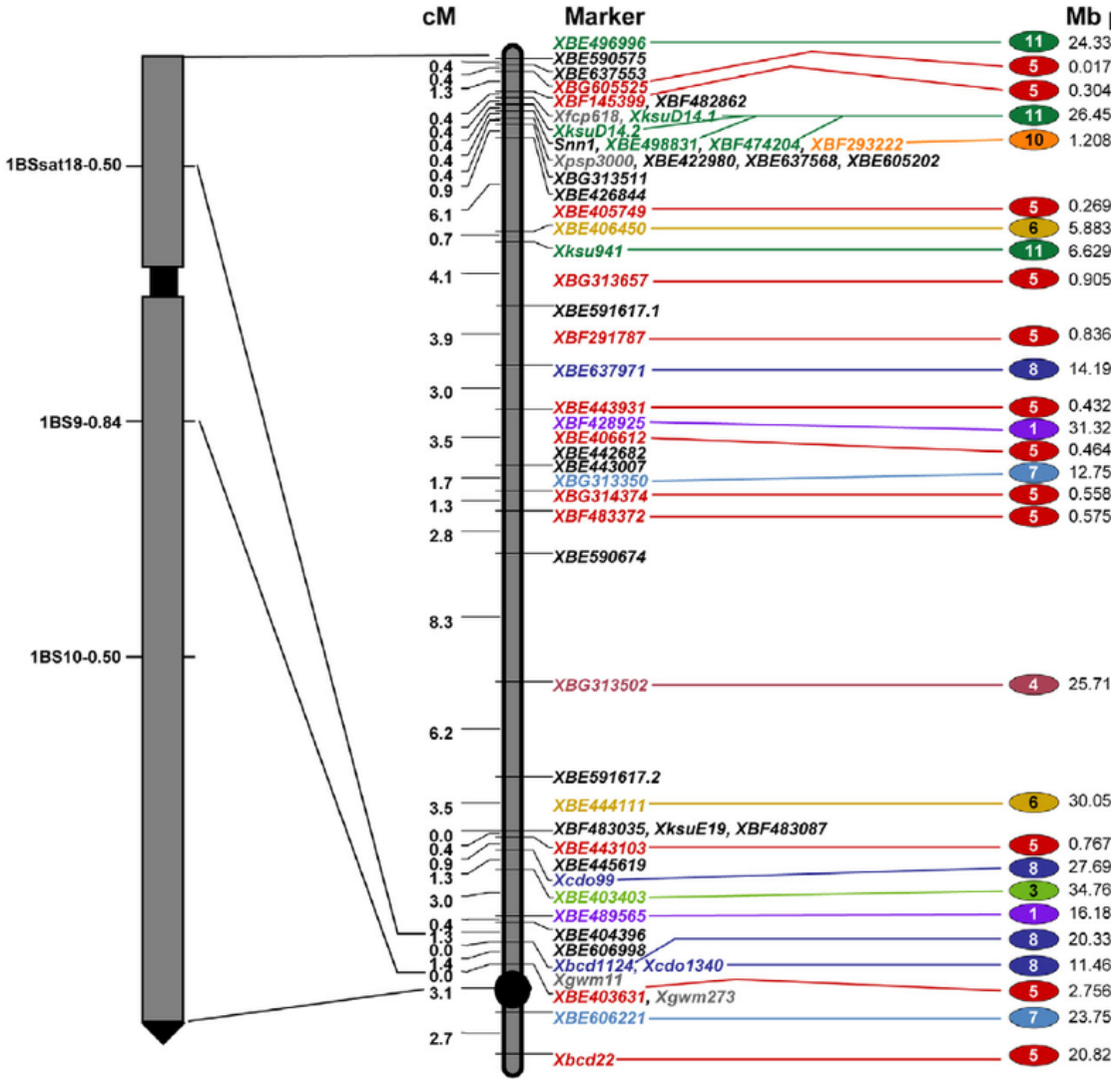


Figure 4.8 Physical (in Mb) and genetic map (in cM) of rice chromosome 1BS (https://acsess.onlinelibrary.wiley.com/doi/full/10.3835/plantgenome2008.03.0181).

The frequencies of recombination between the genes, gene and marker, or markers are used to map the genome. For genetic mapping and the creation of a genetic map, it is necessary to:

a) Create the so-called mapping population (a population of individuals created from different parents and with segregated genes, or traits or properties).
b) Select suitable polymorphic molecular markers (dominant or co-dominant markers, suitable for certain mapping populations).
c) Carry out linkage analysis – calculate recombination frequencies between the genes, gene and marker, or markers, create binding groups, i.e. determine the arrangement of the genes or markers on the same chromosome (determine the LOD score, i.e. determine the logarithm of probability if two loci are linked), determine the order of loci and genes on the linkage map.
d) Locate the genes on chromosomes – FISH (fluorescence in situ hybridization) is currently the fastest method.

In self-pollinated species, i.e. homozygous parents, the mapping population can be created from different types of mapping populations, such as:

a) **F2 population** obtained by self-pollination of the F1 hybrids.
b) **BC population** obtained by backcrossing the F1 hybrids with one of the parents.
c) **RIL population** is a recombinant inbred line obtained from a single selected seed of the F2 generation plant, while this selection is carried out in the next 6-8 generations.
d) **DH population** created by the duplication (dihaploidization) of haploid gametes of the F1 or F2 generation individuals from the haploid pollen grains or embryos.
e) **NIL population** is the close isogenic lines prepared by repeated backcrossing (min. 6x) with a recurrent parent and with subsequent self-pollination, while preparing a homozygous generation for the mapped gene.

4.4 MAS method – marker assisted selection

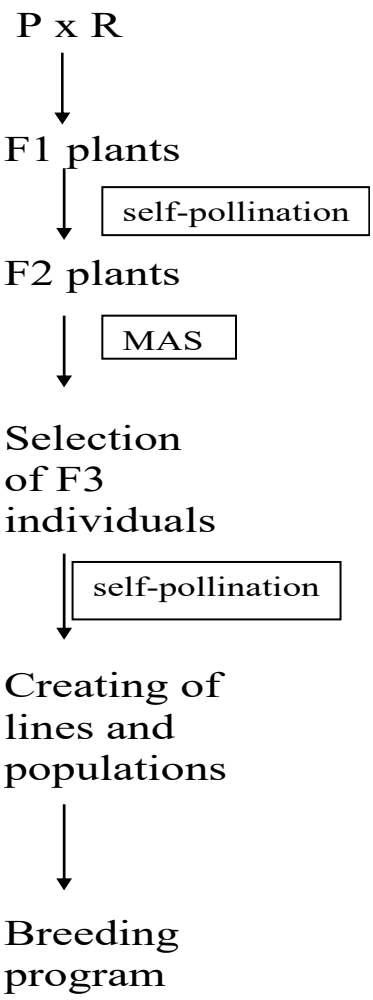
The **MAS** (Marker Assisted Selection) method is based on the fact that the presence of a gene that is closely linked to this gene can be deduced from the detected presence of a marker. If the marker and the gene are located far from each other, the probability that they will be introgressed from parent to generation is reduced by the occurrence of recombinations. Therefore, a necessary precondition for the use of markers in the selection is that they are closely linked to the gene. For this purpose, it is necessary to "saturate" the regions on the linkage genetic map, which means covering the genome with markers so densely that the maximum separation between them is no more than a few cM.

The following criteria are important for marker assisted selection (MAS) in the plant breeding programs:

- The marker should be closely linked to the desired gene (1 cM and less is probably sufficient for MAS)
- An effective molecular method for screening a large population should be developed, which currently is a relatively easy-to-implement method based on the PCR reaction
- The molecular technique used should be reproducible between the laboratories, relatively cheap and easy and effective for the user

Plant material usable for MAS should contain a donor, i.e. a line that contains a significant gene or genes in a homozygous state and markers that can be used to reliably detect these genes in segregating populations.

A codominant marker is more suitable for use in MAS, and it can be used to distinguish a heterozygote from a homozygote. Such DNA markers include STMS, RFLP, SCAR, CAPS and EST-SSR (Table 3.2, 3.3). Dominant markers, such as STS, RAPD or AFLP, are also used in MAS. When using dominant markers, it is advantageous to use two dominant markers that flank the target gene and are mutually in the linkage phase. The conversion of dominant RAPD, ISSR and AFLP markers to codominant markers, such as SCAR and CAPS, is the most advantageous, but time-consuming and often costly, however, the resulting converted markers are highly reliable and laboratory-friendly for screening a large population of plants in MAS.



The MAS method can be very useful in the cumulation of economically important genes, e.g. disease resistance genes, etc. (pyramiding resistance genes) in one and the same plant (Figure 4.9). The use of selected molecular markers linked to important genes is very advantageous when collecting the lines and varieties of plants resistant against pathogens. The donors that contain a certain gene are crossed with selected acceptors to produce the F1 generation. Its self-pollination further creates the F2 generation, from which the individuals with the detected resistance gene in the homozygous state are selected. The presence of a gene is indicated by a molecular marker that is closely linked to the given gene. After self-pollination, the F3 generation is prepared. By combining classical crossing and selection using a molecular marker, the F3 lines with the important gene are obtained.

Figure 4.9 Principle of MAS method

DOI: https://doi.org/10.15414/2023.9788055226897

The following methods are effective in the search for molecular markers: RAPD, RFLP, AFLP and microsatellites. With their help, it is possible to determine which of the loci lies near the gene using the following strategies:

1. NIL (Nearly Isogenic Lines)
2. BSA (Bulked Segregant Analysis)
3. QTL (Quantitative Trait Loci)
4. Molecular marker identification using computer databases on DNA sequences and data mapping

4.4.1 NILs (Nearly Isogenic Lines)

- NILs are created by crossing a sensitive parent with the parent carrying a significant gene (Figure 4.10),
- The progeny is then repeatedly backcrossed with the sensitive parent until the genomes are almost identical except for a small target segment around the specific gene,
- The DNA marker is identified near the inserted gene and this DNA marker will only be present in the individual if the inserted gene is also present,
- This method is limited by the polymorphic ratio between the lines,
- Markers can be obtained if the gene is inserted from related species. In this case, recombinations will be suppressed in heterozygous genotypes with a foreign chromosomal segment.

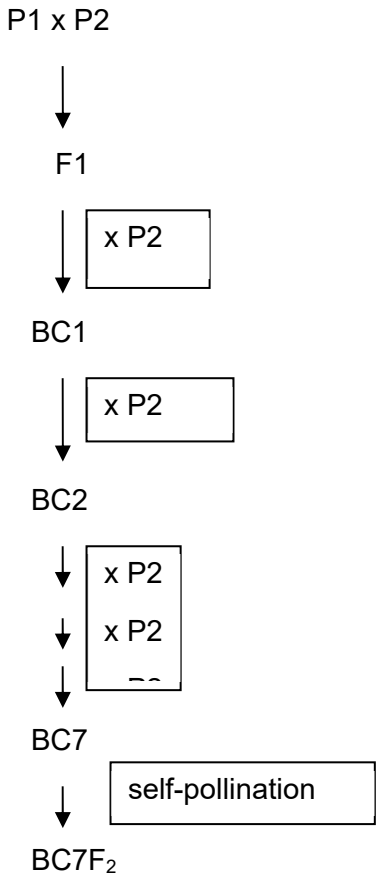


Figure 4.10 Principle of NIL method

4.4.2 BSA (Bulked Segregant Analysis)

- Samples of resistant and sensitive plants from the segregating population in the F2 generation are divided into groups (bulks).
- Molecular markers (RFLP, RAPD and others) are used to compare these groups, which allow the identification of markers linked to a significant gene.
- Group size is based on the frequency with which the unlinked loci can be detected between the groups and on the maximum required distance between marker and gene.
- The higher the number of plants in a given group, the more accurate the results.
- The linkage between the marker and the gene must always be confirmed by an analysis of the segregating population.

DOI: https://doi.org/10.15414/2023.9788055226897

4.4.3 QTL (Quantitative Trait Loci)

- Many important agronomic traits, such as yield, quality, maturity and resistance to biotic and abiotic factors, are controlled by a relatively large number of loci, each of which slightly contributes in a positive or negative way to the final phenotypic value of the trait.
- One quantitative trait is conditioned by a greater number of polygenes, genes of small effect (minor gene).
- The final phenotypic expression is determined by genetic variability in a large number of loci, modified by environmental factors.
- With the help of molecular markers, it is possible to determine the positions of individual QTLs on the chromosomes, the types and size of the gene effects of individual QTLs, and determine which parent carries a positive allele in each QTL.
- The ability to find a linkage between QTL and a molecular marker depends on the size of the QTL effect on the trait, the size of the monitored population and the recombination frequency between the QTL and the marker.

4.4.4 Utilization of the MAS method

- Resistant breeding – acceleration of the incorporation of discrete resistance genes.
- Design breeding, i.e. designing better genotypes under in silico conditions and also physically.
- Pyramiding of major/minor genes into varieties to ensure a more permanent resistance (molecular breeding).
- Improvement of quality parameters.
- Identification of characters that are difficult or financially demanding to identify (male, female fertility genes, etc.).
- Identification of abiotic resistance – drought, water stress, etc.
- Acceleration of backcrossing using DNA markers.
- Using DNA markers, mapping of agronomically significant traits, such as resistance to pathogens – diseases, pests, tolerance to abiotic stresses, qualitative and quantitative traits.

5 GENOMICS AND BIOINFORMATICS

The term genome was derived by the German botanist Hans Winkler in 1920, from the words gene and chromosome, which he called a haploid set of chromosomes. The term *genomics* was first used by the American geneticist Thomas Roderick in 1986. Subsequently, in 1987, the first issue of the journal *Genomics* was published, devoted to this newly emerging field of science.

The genome is the total genetic material of a cell or individual in the haploid set of chromosoms (Figure 5.1). In a plant cell, there are genomes of the cell nucleus, chloroplasts and mitochondria. Animals have genomes of the cell nucleus and mitochondria. In the bacterial and archaeal cell, the genome is made up of the chromosome and plasmids. The genome of viruses is made up of DNA or RNA virion molecule(s). The bacterial genome, consisting of a circular chromosome, contains about 0.01 pg (1 picogram = 1012 g) of DNA, the haploid mammalian genome 3-6 pg of DNA, but some amphibians and plants can contain more than 100 pg of DNA in the haploid genome.

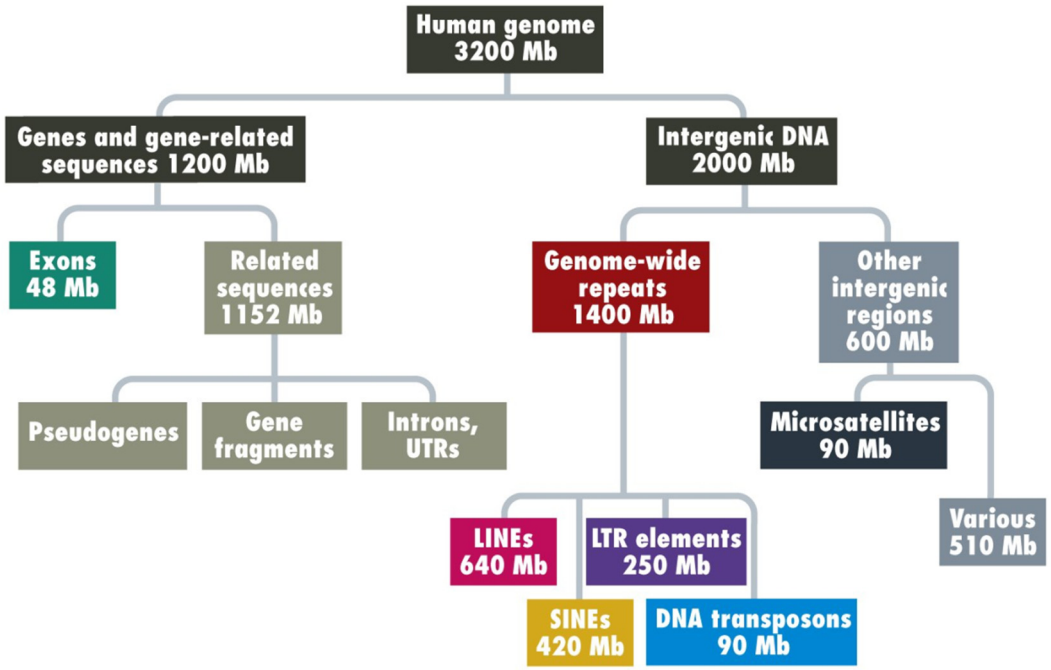


Figure 5.1 The composition of the human genome. LTR – long terminal repeat, LINE – long interspersed nuclear elements, SINE – short interspersed nuclear elements. (Brown, 2007, modified).

DOI: https://doi.org/10.15414/2023.9788055226897

The genomes of organisms are characterized by different sizes, structures and ploidy. This variation can exist even within a single species (Figure 5.2). From this point of view, the relevance of genomics and the consequently derived 'omics' sciences is irreplaceable.

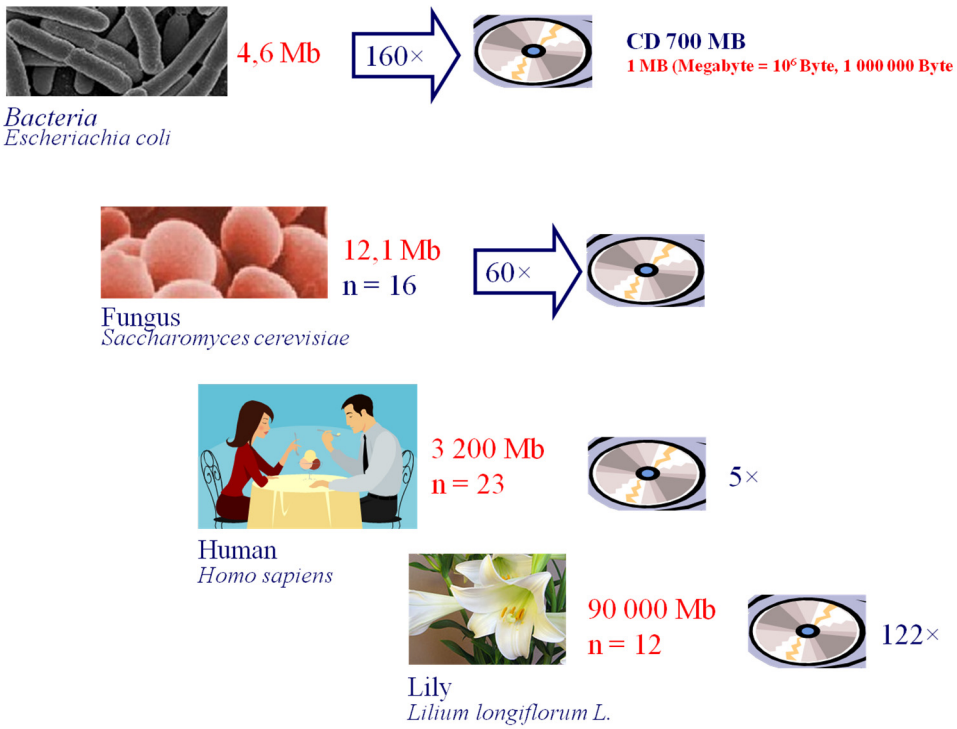


Figure 5.2 The comparison of genome (Mb) size and CD capacity. (© Genova 2005).

The genome size is the total amount of DNA in single copy of genome. The amount of DNA in gamete is defined as DNA content (1C value), independent from organisms ploidy. The amount of nucleus DNA is expressed by amount of DNA in picograms (pg) or number of Mbp of DNA for 1C value. The nuclear DNA content as C-value varies widely in higher plants from 0.15 to over 200 pg (Table 5.1). Even the smallest plant genome is larger than that found in *Drosophila melanogaster* (165 000 kb). The amount of DNA has also been shown to be related to the geographical distribution of crops, phenology, biomass and the sensitivity of growth to environmental conditions (temperature and frost). Plants with a higher DNA content and a specific chromosome structure have been found to be more resistant to radiation damage (*Ginkgo biloba* L.).

DOI: https://doi.org/10.15414/2023.9788055226897

Table 5.1 Examples of DNA content in plants (http://www.rbgkew.org.uk/cval/homepage.html)

Genus	Species	1 C (pg)
Arabidopsis	*thaliana*	0.125
Oryza	*sativa*	0.5
Zea	*mays*	2.73
Nicotiana	*tabacum*	5.85
Allium	*sativum*	16.23
Ginkgo	*biloba*	9.95
Pinus	*poderosa*	24.2

5.1 Regions of the genome of the plant cell nucleus

In eukaryotic cells, most of the DNA is located in the nucleus. The nucleus is the site of DNA synthesis and RNA production in the process of transcription. Nuclear DNA forms complexes with proteins and, together with other proteins, forms chromosomes - large nucleoprotein complexes. Each chromosome contains one linear DNA molecule. The nuclei of different species contain different numbers of chromosomes, and in each species the chromosomes vary in length. Other cell organelles, such as mitochondria and chloroplasts, contain their own DNA in the form of closed circular molecules, without histone proteins, as in the case of bacterial and archaeal cells.

The nucleus of a plant cell contains linear DNA molecules that are organized in gene pools called chromosomes. Higher plants are mostly diploid, that is, they have two sets of chromosomes in each cell nucleus, with each set of chromosomes having relatively identical genes. The set of chromosomes in the nucleus of the cell with the genes make up the genome of the nucleus.

There is no direct relationship between genome size and chromosome number. For example, the genome of wheat (*Triticum* sp.) has seven chromosomes, but the amount of DNA is ten times greater than in the genome of the potato lily (*Solanum tuberosum* L.) (12 chromosomes). This relationship is referred to as the „C-value paradox“ (C is the DNA content for the unreplicated genome). Such a relationship occurs not only between different species but also between species within genera.

In addition to genes, the plant nuclear genome contains intergenic deoxyribonucleic acids (Table 5.2). Within the intergenic DNA are stretches with separate arrangements of repeating nucleotide sequences in the DNA molecule. While genes have a relatively stable position in the genome of the cell nucleus, certain stretches of intergenic DNA, the mobile genetic particles, can change their position in the genome.

DOI: https://doi.org/10.15414/2023.9788055226897

Table 5.2 Overview of the structure of the plant nuclear genome.

<table>
<tr><td rowspan="7">Plant nuclear genome</td><td rowspan="3">Genes and gene-like sequences of nucleotides in a DNA molecule</td><td colspan="2">Genes</td></tr>
<tr><td rowspan="3">Gene-like sequences of nucleotides in a DNA molecule</td><td>Pseudogenes</td></tr>
<tr><td>Gene fragments</td></tr>
<tr><td>Introns</td></tr>
<tr><td rowspan="4">Intergenic DNA</td><td rowspan="3">Separately ordered repeating sequences of nucleotides in a DNA molecule</td><td>Moving genetic particles
- DNA transposons,
- Retrotransposons with LTR,
- Retrotransposons without LTR,
(retroposons - SINE, LINE)</td></tr>
<tr><td>Simple repeating sequences of nucleotides in a DNA molecule</td></tr>
<tr><td>Duplications of DNA sections</td></tr>
<tr><td rowspan="3">Consecutively arranged repeating sequences of nucleotides in a DNA molecule</td><td>Satellite DNA</td></tr>
<tr><td>Minisatellite DNA</td></tr>
<tr><td>Microsatellite DNA</td></tr>
</table>

Note: LTR – long terminal repeat, LINE – long interspersed nuclear elements, SINE – short interspersed nuclear elements.

5.1.1 Transposable elements of the plant genome

Transposable elements were first discovered in plants as elements with significant influence on genome structure and gene activity. They have the ability to change the structure of genomes and individual genes and control their expression through translocation, insertion and excision of sequences, chromosome breaks and ectopic recombination. By incorporating them directly into genes or into their surrounding sequences, they can affect gene expression or inactivate them completely. They are involved in somatic instability, sterility, germination disorders and inherited diseases. The study and understanding of transposable elements of the genome and their mechanisms of action is also a pathway to understanding plant genes and genome evolution.

Transposable elements are defined by the mechanism by which they propagate in the genomes of organisms. The mechanism of replication involving the production of mRNA and subsequently cDNA by the reverse transcription enzyme (reverse transcriptase) is characteristic of class I transposable elements, retrotransposons (Figure 5.3). The second class of transposable elements, transposons, uses a cut-and-paste mechanism at the DNA level. This class includes, for example, the Ac elements mainly distributed in the human genome.

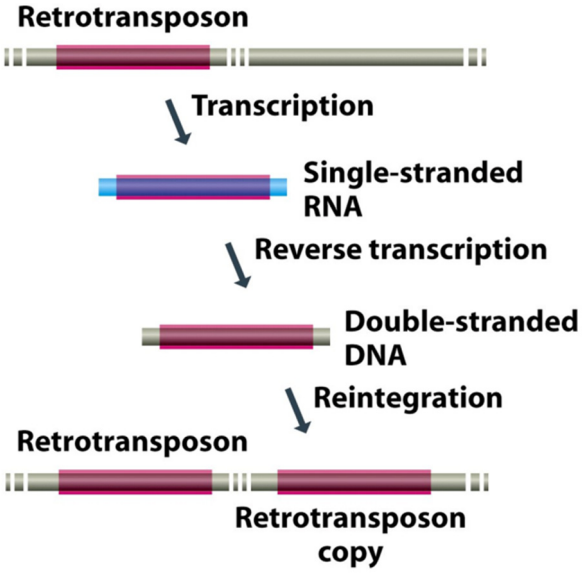


Figure 5.3 Retransposition of retrotransposons. (Brown, 2007, modified).

In many cases, retrotransposons can make up more than 50% of the genome of the cell nucleus. Several types of retrotransposons are highly abundant in the euchromatin regions of chromosomes. Mutations caused by incorporation of retrotransposons into or near a gene can result in inactivation or altered gene expression and ultimately the structure of the protein. Most of the transposable elements of plant genomes are either present in an inactive form or their expression is repressed.

The properties of retrotransposons are as follows:
(a) mutations caused by the incorporation of a retrotransposon into the genome are stable because their mechanism of translocation in the genome is based on the principle of replication,
(b) the site of retrotransposon translocation is different from the site of the original copy, which may condition the occurrence of random mutations,
c) retrotransposon activity can be induced by stress conditions (abiotic and biotic conditions),
(d) due to their preferential incorporation into genes or regions of genes, they have a significant potential for inducing mutations,
(e) the low copy number facilitates the identification of the retrotransposon insertion responsible for a specific mutation,
(f) retrotransposons are active in many plant species.

Retrotransposons are divided into two subclasses based on their structure and transposition cycle:

- LTR retrotransposons,
- retrotransposons without LTR regions.

LTR retrotransposons are formed by long terminal repeats (LTRs) of nucleotides located at the 5′ and 3′ ends. LTR retrotransposons are further subdivided into two categories according to nucleotide sequence similarity and organization, namely the Ty1-*copia* and *gypsy* categories. The *Ty* designation represents the *yeast transposon* designation, *copia* – a copy of the retrotransposon of the common *Drosophila* (or Pseudoviridae*)* and *gypsy* – a retrotransposon of the common *Drosophila (*Metaviddidae*)*.

DOI: https://doi.org/10.15414/2023.9788055226897

Copia-like retrotransposons (Figure 5.4) have an internal domain (the portion between the 5′ LTR and the 3′ LTR) formed by open reading frames (ORFs) for the envelope protein (GAG), aspartic protease (AP), integrase (IN), reverse transcriptase (RT), and RNase H (RH). The envelope protein of retrotransposons is one of the common features shared by retrotransposons and retroviruses. Aspartic protease cleaves the active polyprotein into functional components, integrase mediates the incorporation of the cDNA into a new location in the genome, reverse transcriptase (reverse transcription enzyme) is responsible for the formation of cDNA copies, and RNase H is important for retrotransposon replication.

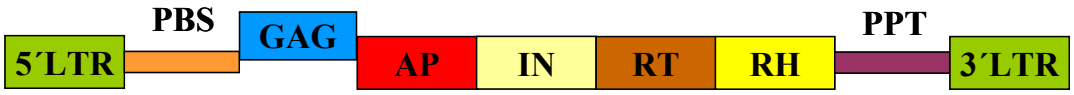


Figure 5.4 Schematic of the structure of copia-like retrotransposons. 5′ LTR and 3′ LTR – Long Terminal Repeats (LTR), long nucleotide repeat sequences at the 5′ and 3′ ends, Primer Binding Site (PBS), Capsid Protein gene (GAG), AP – Aspartic Protease gene, IN – Integrase gene, RT – Reverse Transcriptase gene, RH – RNase H gene, PPT – Polypurine Tract. (Vicient, et al. 2001, modified)

Gypsy-like retrotransposons have a different ORF order compared to copia-like retrotransposons (Figure 5.5). The 5′ LTR contains the nucleotide order of the promoter, whereas the 3′ LTR contains the order of the terminator and polyadenylation end. The 5′ LTR includes regulatory tandem nucleotide sequences, cis, which are located in the opposite direction of the transcription start site (U3 region) or in the direction of the UTR (Untranslated Region, regions where no transcription of genetic information occurs).

Figure 5.5 Schematic of the structure of gypsy-like retrotransposons. The description of the individual structures is identical to Figure 5.4.

The subclass of retrotransposons without LTR regions is represented by LINE (Long Interspersed Elements) and SINE elements (Short Interspersed Elements). The internal domain is similar to the domain of LTR retrotransposons.

5.1.2 Repetitive nucleotide sequences of the plant genome

Repetitive DNA consists of sequences that are present in more than one copy in each haploid genome. Repetitive DNA can be divided into two general types: Moderately repetitive DNA consists of relatively short sequences that are repeated typically 10 to 1,000 times in the genome. For examples, genes for tRNAs and rRNAs. Highly repetitive DNA consists of very short sequences (typically fewer than 100 base pairs) that are present many thousands of times in the genome, often organized as long regions of tandem repeats.

Another important source of DNA polymorphism is satellite DNA, which is divided into satellite DNA (100-5000 bp), minisatellite DNA (10-20 bp) and microsatellite DNA based on the size of the repeated nucleotide sequences, which is based on repeats of a smaller number of

DOI: https://doi.org/10.15414/2023.9788055226897

nucleotides (di-, tri- and tetranucleotides), with the number of repeats being variable within organisms. Mini and microsatellite DNA sequences represent a significant source of polymorphism in eukaryotic genomes and are also very useful for genotype analysis and map construction. Satellite DNA refers to tandemly repeated DNA elements that were first isolated by centrifugation in density gradients from satellite sections. Satellite DNA forms a substantial part of the large genome complexes of eukarya. Some of the satellite DNA elements are nearly universal across species. It is known that repetitive sequences of satellite DNA are mainly part of heterochromatin and are located mainly in centromere and telomere regions. Analyses confirm that these elements are structural and functional components of the chromosomes of eukarya.

More than 160 different groups of repetitive satellite DNA sequences have been reported and described in the genomes of higher plants. The most frequently occurring repetitive sequences in the plant genome are dinucleotides (AC)n and (GA)n. The most frequently occurring repetitive trinucleotides are (AAG)n and (AAT)n.

This type of sequences is divided into three categories:
1. complete repeat sequences (perfect repeat sequences) without interruptions,
2. imperfect repeat sequences with one or more interruptions,
3. compound repeat sequences with adjacent tandem simple repeats of different sequences.

5.1.3 The chloroplast genome

Chloroplasts contain DNA referred to as chloroplast DNA (cpDNA). In higher plants, the size of cpDNA ranges from 120 to 160 kb, in algae from 85 to 292 kb. Chloroplast DNA takes the form of a closed circular molecule. In some species, especially with a large cpDNA molecule, it is not impossible that this molecule exists in a linear form.

The number of cpDNA molecules in a cell depends on two factors: the number of chloroplasts and the number of cpDNA molecules in each chloroplast. All cpDNA molecules contain essentially the same set of genes, but these genes can be arranged in different ways in different species. The basic gene assembly includes genes encoding ribosomal RNAs, transfer RNAs, some ribosomal proteins, various polypeptide components of photosystems involved in solar energy fixation, a catalytically active subunit of the enzyme ribulose-1,5-bisphosphate carboxylase, and four subunits of chloroplast-specific RNA polymerase. For example, *Nicotiana tabacum* cpDNA contains 155 844 bp and 150 genes.

Most cpDNA molecules include several large inverted repeats that contain genes for ribosomal RNA. These repeats are 10 to 76 kb in length and are localized differently in different cpDNA molecules. The development of functional chloroplasts depends on the expression of both nuclear and chloroplast genes. The products of nuclear genes that are functional in chloroplasts must be imported into chloroplasts from the cytosol. Thus, the action of these proteins in the chloroplast must be coordinated with the proteins encoded in the cpDNA. Functional chloroplasts are thus dependent on the coordinated activity of both nuclear and chloroplast gene products.

5.1.4 The mitochondrial genome

The size of the mitochondrial DNA (mtDNA) molecule is extremely variable, ranging from 6 kb in the malaria-causing parasite *Plasmodium* to 2500 kb in some higher plants. Each mitochondrion contains several copies of DNA, and since each cell contains several mitochondria, the number of mtDNA molecules in a single cell can be very high.

Most mtDNA molecules are circular in shape, but in some species, such as the alga *Chlamydomonas reinbardtii*, they are linear in shape. The genetic organization of these molecules can be highly variable. The mtDNA of plants is much larger compared to other organisms. And it also has a more variable structure. The mtDNA molecules of higher plants contain many noncoding sequences, including several duplicated ones. Physical mapping of plant mitochondrial genes has shown that circular molecules are localized at different sites in different species, and even in related species. This implies that the mtDNA of higher plants has undergone many genetic rearrangements.

Most mtDNA gene products are functional only within the mitochondria, but do not function here independently. Nuclear gene products are imported into the mitochondria that have the ability to amplify or facilitate mitochondrial function. For example, mitochondrial ribosomes are assembled from ribosomal RNA transcribed from genes in the mitochondria and from ribosomal proteins encoded by nuclear genes. Ribosomal proteins are synthesized in the cytosol and imported into the mitochondria where ribosome assembly occurs.

5.2 Genomics

Genomics is a field of study in biology focusing on the knowledge of all genes in the genome with respect to the complexity of the structure and function of the genome, the activity of genes in the genome, the reasons for their activity and the location of their activity in the cell and the organism, and the interplay of genes in the genome in the formation of the organism.

Genetics is concerned with the complex but more or less linear relationship between the structure and function of a gene, which can be expressed by the relationship DNA (gene) – RNA – protein (trait). Genomics studies the complexity of the structure and function of genes in the genome. Bioinformatics uses computer science to work with biological (genomic) data.

Genomics is aimed at understanding the DNA sequences of organisms and non-cellular biological systems, genetic mapping, creating genome surveys, searching for genes and other functional components of genomes, determining the significance of DNA sequences, and comparing the genomes of different organisms. In genomics, methods of molecular biology and bioinformatics are applied.

Genomics is an interdisciplinary field of science studying the structure, function and evolution of the genome of organisms. Genomics is a broad research domain that encompasses several specialised and often overlapping areas. The rapid emergence of next generation sequencing (NGS) methods, high-throughput analysis of genomes, and bioinformatics methods provide a sound methodological and infrastructural basis for this dynamic field of science.

The methodological approach of genomics uses the complete genome sequence, on the basis of which it examines genes using the reverse genetics approach, identifying their function in the genome, the interactions of genes with each other, as well as interactions with the environment. In this approach, a mutation is targeted in a known gene sequence and then the phenotypic variation resulting from the change is examined. Conventional genetics, also referred to as forward genetics, studies biological objects from their functions to the genes that determine them.

Genome mapping is the creation of a genome description sheet. The order of DNA nucleotides in the genome is determined and genome maps are made. The structure and function of each gene as well as the effect of gene action on the genes in the genome of the organism is determined. Intragenic (alleles, loci) and intergenic (gene-gene interactions) relationships of genes in the genome are recognized.

Genomics is carried out in three stages – (a) determination of the nucleotide sequence of the genome, genome mapping (b) locating the gene and determining its function, (c) applying the knowledge to biological systems.

a) Genome mapping: determining the order of DNA nucleotides in the genome and constructing a genome map; determining the structure and function of each gene; understanding the intragenic (alleles, loci) and intergenic (gene interactions) relationships of genes in the genome.

b) Develop an overview of how genes in the genome interact and respond to the environment.

c) Use knowledge of genome structure and function in biological systems.

As new information was acquired, detailed maps and sequences of genomes were created and genomics was divided into structural genomics, studying the structure of the genome, comparative genomics, studying the evolution of the genome, and functional genomics, involving the analysis of the transcriptome and proteome.

A *transcriptome* is the collection of all the mRNAs in a cell, tissue, organ or organism (Figure 5.6). It reflects the expression level of a gene(s) and the stability of the transcribed stretches, transcripts. The transcriptome of a non-cellular biological system is formed only in the cell. Transcriptomics is concerned with monitoring gene/gene expression and detecting differences in gene expression depending on the internal conditions in the organism, the external conditions in which the organism, the stage of development of the organism, the cell, tissue or organ of the organism.

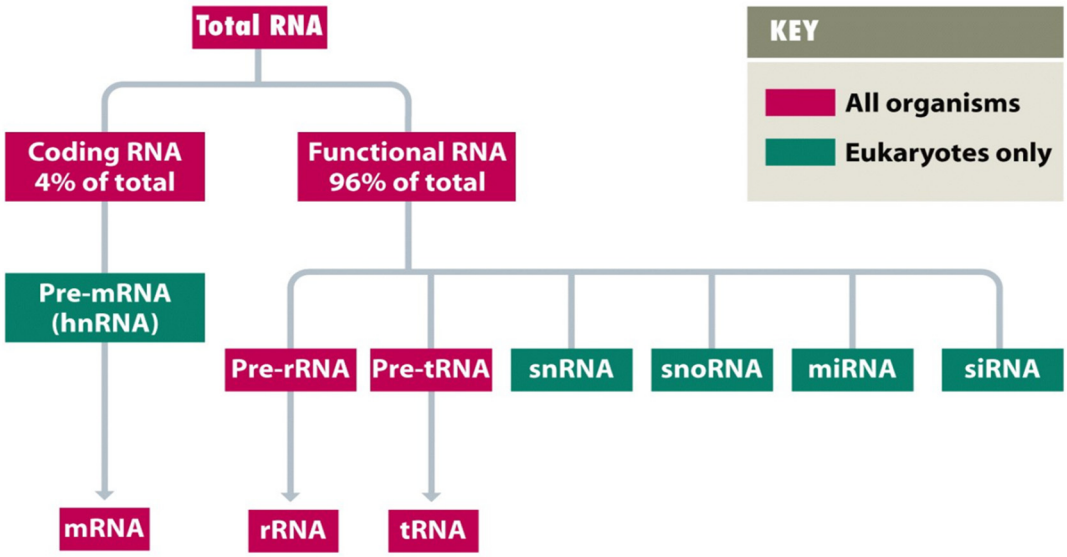


Figure 5.6 The RNA content of a cell. The types of RNA present in all organisms and those categories found only in eukaryotic cells. (Brown, 2007, modified). rRNA – ribosomal RNA, tRNA – transfer RNA, snRNA – small nuclear RNA, snoRNA – small nucleolar RNA, miRNA – microRNA, siRNA – short interfering RNA, pre – precursor.

A *proteome* is a collection of all the proteins of a particular system, such as the organelles of a cell, cell, mesh, organ, or organism. All functionally distinct forms of proteins are part of proteome. Functionally distinct forms are created by post-translational modification of a protein encoded by a single gene. The proteome of a cell changes dynamically depending on the cell cycle, cell development, the cell's response to changes in metabolism, and changes in the cell's environment (Figure 5.7).

DOI: https://doi.org/10.15414/2023.9788055226897

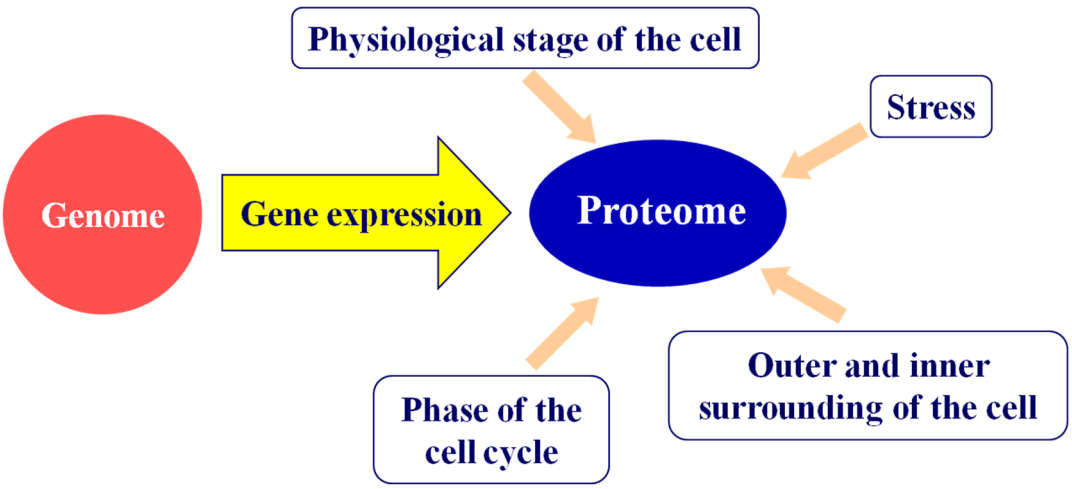


Figure 5.7 Proteome of the cell is dynamically changed depending on the cell cycle, cell development, cell response on the metabolism changes and changes around the cell.

Genomics is divided into structural (structure of genes and whole genomes in cells, cell organelles and non-cellular biological systems), functional (searching for genes and determining their function), comparative (comparing genomes of organisms of different taxa), population (comparison of genomes of organisms within a population), computational (processing of genomic data by bioinformatic methods), personal (knowledge of the primary structure, DNA sequence of the genomes of the cell nucleus and mitochondria of a human individual, its analysis and estimation of the genetic predisposition of a person).

Structural genomics is a branch of genomics focused on understanding the primary nucleic acid structure of genomes and genes in the genomes of a cell, organism, or non-cellular biological system.

Comparative genomics focuses on the differences, or similarities, between the genomes of individuals of different species or of the same species. Each individual has its own version of the genome. Comparative genomics applies the available sequences of model organisms and another type of genome to extract new relevant information by studying their genomes. Using the tools of comparative genomics, it is possible to identify previously unannotated genome sequences and subsequently confirm their existence by molecular analyses.

Single nucleotide polymorphisms (SNPs) cause variation in the genomes of individuals. A SNP is a nucleotide difference at a particular position in the genome of individuals (1st person ATTCCTA, 2nd person AGTCCTA, 3rd person AATCCTA). More than 1.4 million SNPs have been identified in the human genome (1 SNP in approximately 2 kb of nucleotide sequence). SNPs are in the non-coding part of the genome, but ~60,000 are in genes. SNPs in genes influence their action. SNPs confirm the individual biological characteristics of each person.

Proteins often have structural units or domains, conserved in different species of organisms. Genome-wide analyses indicate the presence of a large number of evolutionarily conserved sequences in the genome of humans and several model organisms. By comparing nucleotide sequences of genomes, changes responsible for species divergence have been noted. Comparative genomics is a powerful tool for studying the genetic relatedness of species. Phylogenetic cluster dendrograms can be constructed from DNA sequences. Comparative genomics can be applied in different ways.

DOI: https://doi.org/10.15414/2023.9788055226897

Functional genomics focuses on the study of the function of gene sequences and their expression under specific conditions. The key task of functional genomics is to identify the interrelationships of genes in the genome and the interaction of their activity with the environment. Functional genomics involves the analysis of the transcriptome, the set of RNA molecules transcribed by a particular genome, and the proteome, the set of proteins encoded by a particular genome.

Personal genomics is a branch of genomics focused on the description and analysis of an individual's genome using bioinformatics tools. Personal genomics is close to population genetics. Genome description (genotyping) is most often by single nucleotide polymorphism (SNP) or partial or complete determination of the nucleotide sequence of the genome. Based on the knowledge of the genome, bioinformatics analyses are used to compare the genomes of individuals and to detect associations between diseases and genes or loci.

Pharmacogenomics is a branch of genomics concerned with the relationship of drug action at the whole genome or transcriptome level. SNPs (single nucleotide polymorphisms) are prospective for pharmacogenomics. SNPs clustered in blocks form a haplotype. Due to the small number of recombinations, most sections of human chromosomes can be characterized by several haplotypes. Based on haplotype mapping, it is possible to identify risk haplotypes for a particular drug or group of drugs by applying SNP DNA chips.

Metagenomics (Environmental genomics, ecogenomics, community genomics) is the genomics of organisms living in their natural environment. It has been found by genomic analysis that there are 5,000 different viruses in 200 litres of seawater. Subsequent analyses confirmed the presence of more than 1 000 virus species in human faeces and the possibility of up to a million different viruses, including bacteriophages, per kilogram of marine sediment. Metagenomic studies in the Sargasso Sea have confirmed nearly 200 new species of organisms, including 148 previously unknown bacteria.

The perspectives of genomics are in the control of the synthesis and formation of the three-dimensional structure of proteins, the construction of synthetic life forms, the control of ontogeny, the diagnosis of hereditary stresses, gene therapy and the reconstruction of the history of life on Earth.

5.2.1 Genomics methods. Genome mapping.

Genetic mapping is the process of identifying and locating coding or non-coding regions of DNA or RNA in a chromosome. Genetic mapping results in gene and genome maps. A genome map shows the arrangement of genes or other markers and the distances between them in each chromosome (Figure 5.8). Genome maps are (a) genetic linkage maps and (b) real maps. Genetic linkage maps show the relative locations of genes or specific DNA markers in a chromosome. Genetic linkage maps can depict: coding regions (genes), phenotypic traits and genetic markers.

Real maps show the chemical, physical or functional properties of a DNA molecule. They are based on mapping the actual position of genes, or genetic markers, in a chromosome. Real maps include: chromosome maps, cDNA maps, cluster maps, macrorestriction maps and sequence maps. In a chromosome map genes or specific stretches of DNA are localized in a particular chromosome. Hybridisation of the labelled probe with a complementary strand of DNA identifies a specific stretch in the chromosome. cDNA maps show the location of the functional gene in the chromosome.

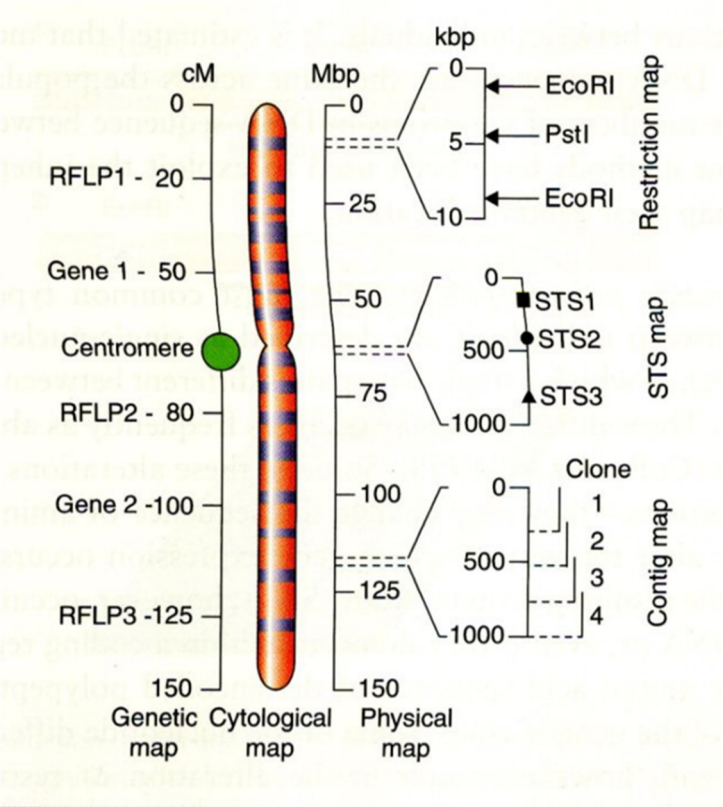


Figure 5.8 The different types of cytological, genetic and physical map of a chromosome. Genetic map distances are based on crossover frequencies and are measured in centiMorgans (cM), while physical distances are measured in megabase pairs (Mpb) or kilobase pair (kbp). (Snustad and Simmons, 2012).

5.2.2 Determination of the nucleotide sequence of the genome

DNA nucleotide sequencing is a technique for determining the exact order of nucleotides in a DNA molecule. Short stretches of DNA cut by restriction enzymes are isolated by gel electrophoresis. The two most commonly used methods for determining the nucleotide sequence are (a) the chemical cleavage method developed by Maxam and Gilbert at Harvard University in 1977, and (b) the chain termination method developed by Sanger, Nicklen and Coulson at Cambridge University in 1977.

Maxam and Gilbert method

In the chemical method (Maxam and Gilbert), the original DNA molecule is chemically degraded. The 5′ or 3′ ends of single-stranded DNA are radiolabeled (32P) using polynucleotide kinase (5′ end) or terminal transferase (3′ end). In addition to single-stranded DNA, restriction fragments of DNA can also be used, produced by cleavage of the restriction enzyme, which removes a small portion from one end of the DNA molecule, leaving a fragment with only one radiolabelled end (Figure 5.9). Double-stranded DNA can only be used if one of the ends of the molecule is radiolabelled. After chemical modification of the nucleotide, the DNA strand is cleaved by a reaction in which the sugar-phosphate skeleton of the DNA strand is fragmented.

Chemical modification of purine bases (A and G): the DNA molecule is exposed to an acid followed by dimethyl sulphate, which causes methylation of A at the third position and G at the seventh position of the sugar component of the nucleotide (deoxyribose). Subsequent

reaction of the DNA with an alkali and piperidine cleaves the polynucleotide chain before the purine residue.

Chemical modification of pyrimidine bases (C and T): the DNA molecule is exposed to hydrazine, which causes hydrolysis of T. The action of piperidine cleaves the chain before the pyrimidine residue. If the reaction is carried out in the presence of 1M or 2M NaCl, hydrolysis of C occurs.

These chemical reactions result in radioactively labelled DNA molecules that are extended from a common point (the radioactively labelled end) to the point where the chemical cleavage occurred. The cleavage products are separated by electrophoresis. Comparison of the electrophoreograms determines which base was present and in what order in the nucleotide chain.

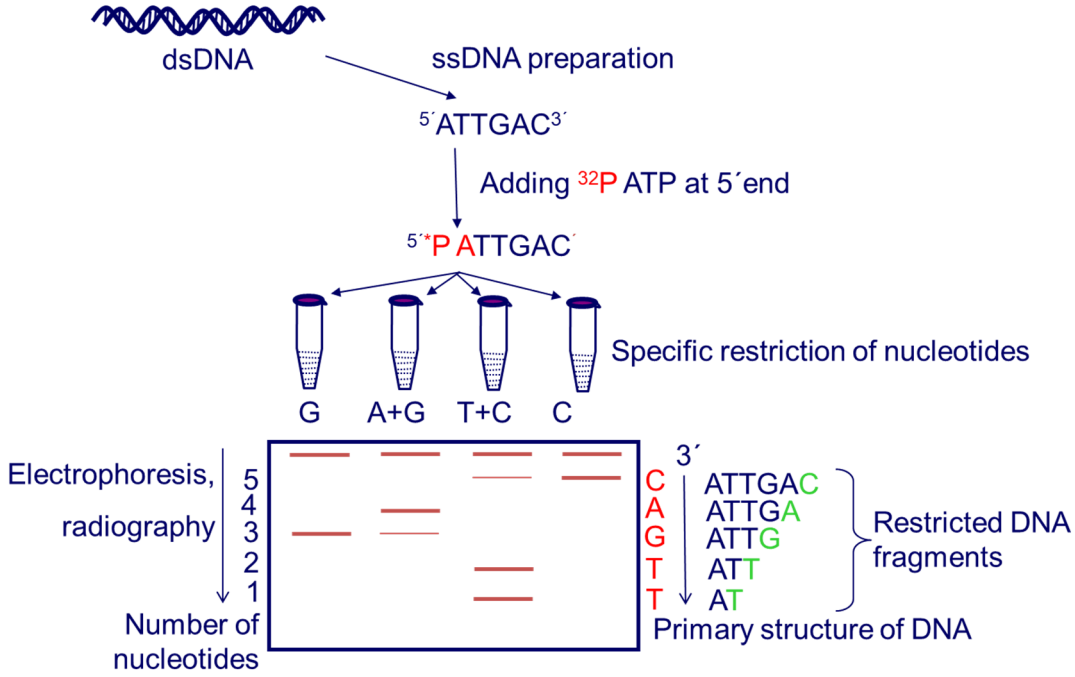


Figure 5.9 Determination of primary structure DNA by Maxam, Gilbert method. (© Genova 2005).

The Sanger method

The chain termination method is based on the principle of double-stranded DNA synthesis according to single-stranded DNA in four reaction mixtures that contain, in addition to the necessary DNA synthesizing components, a small portion of different ddNTPs - dideoxynucleoside triphosphates (stop DNA chain elongation), a labeled primer complementary to the end of the strand, and all four nucleotide triphosphates (Figure 5.10). Four different reactions, in which a different dideoxynucleoside triphosphate each time synthesizes DNA strands shorter than normal DNA. Electrophoresis on a polyacrylamide gel and autoradiography determine the nucleotide sequence. 5′, 3′ ddNTPs differ from dNTPs in that they do not have an OH group at the 3′ position of the deoxyribose. They can be incorporated by DNA polymerase into the synthesizing strand via their 5′-triphosphate group. The absence of the 3′-OH group prevents the formation of phosphodiester bonds with the next dNTP, i.e. chain elongation is stopped and completed.

DOI: https://doi.org/10.15414/2023.9788055226897

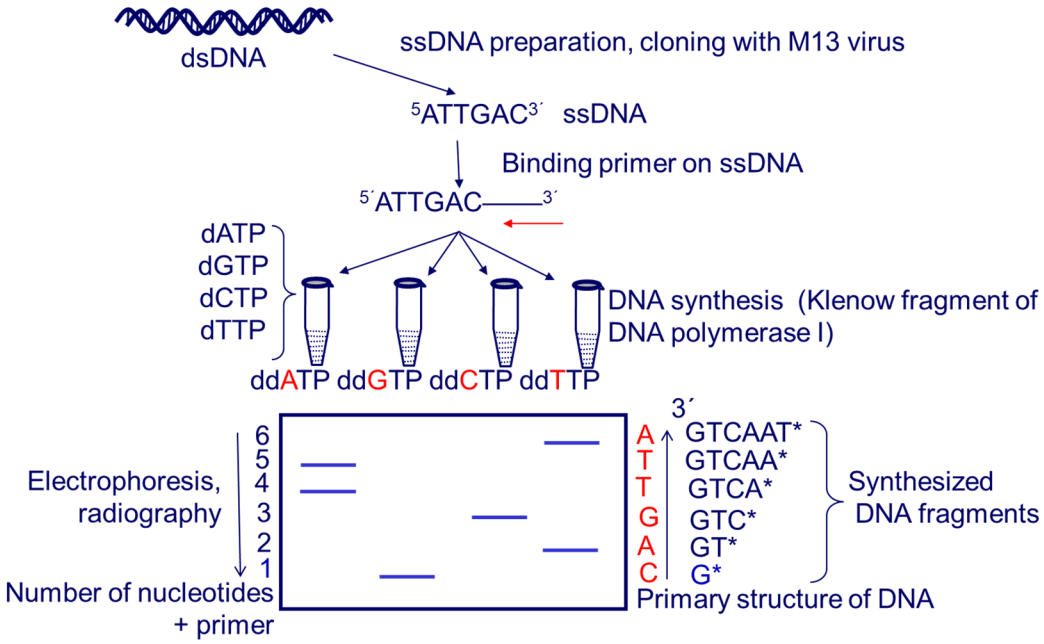


Figure 5.10 Determination of primary structure DNA by Sanger, Coulson method. (© Genova 2005).

Before the actual nucleotide sequence is determined, the single-stranded DNA is cloned into phage M13. The M13 genome is made up of a circular, single-stranded DNA molecule that is complemented into a double-stranded circular molecule inside an *Escherichia coli* bacterial cell (phage M13 with cloned DNA is incorporated into *Escherichia coli*). This double-stranded DNA is the template by which single-stranded DNAs are formed. The synthetic primer binds to the single-stranded DNA above the restriction site that was used to clone the DNA into M13.

The synthesis takes place in four separate reactions. In each reaction there is a different ddNTP in addition to all four dNTPs (dTTP, dCTP, dGTP, dATP). It is of the four dNTPs (usually dATP) and is radiolabeled. Each of the four reactions forms a group of partially synthesized DNA fragments of different lengths that have the same 5' end (determined by the primer), each fragment is radioactively labeled and terminated with a ddNTP.

DNA from these reactions is denatured and size separated on a polyacrylamide gel. The order of the bands on the autoradiograph is determined from the bottom of the gel upwards, which corresponds to the 5′ → 3′ sequences of the strand synthesized in vitro and is thus complementary to the incorporated template in the M13 phage (i.e., all bands in the T-track correspond to the position of the A sequences on the template DNA).

DOI: https://doi.org/10.15414/2023.9788055226897

Automated sequential method

Despite the sophistication of the previous methods, these are time-consuming and labor-intensive for large-scale use. Since these first achievements in sequencing, many sequencing techniques have been developed. DNA sequencing has undergone three generations of major evolution (Figure 5.11). All sequencing technologies have their own advantages and disadvantages.

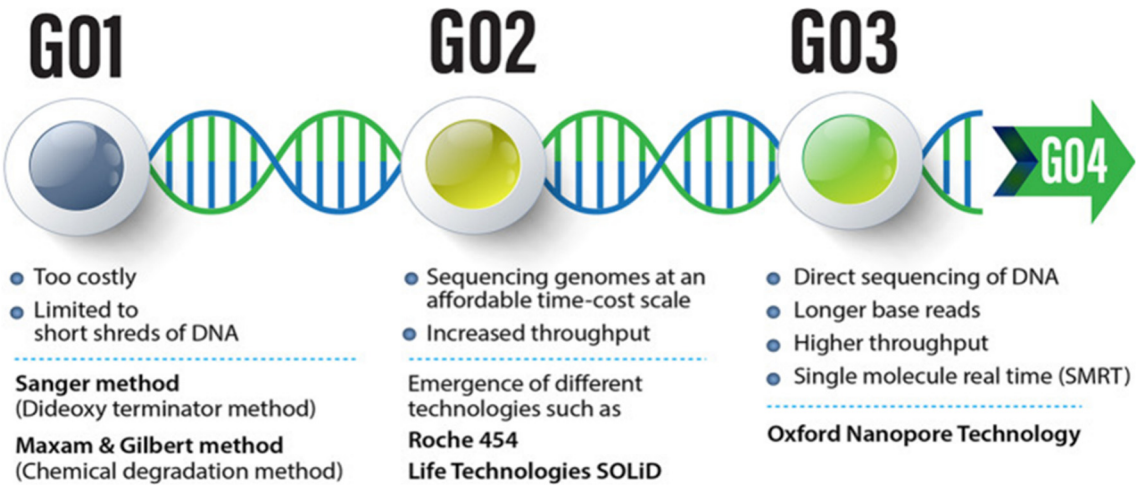


Figure 5.11 A glance at DNA sequencing generations and some features of each generation (Mohammadi et al., 2021).

In 1987, Prober and colleagues developed an efficient system for determining the nucleotide sequence of DNA using fluorescently labeled ddNTPs. Each of the four ddNTPs has incorporated a different fluorochrome; in this respect, four separate reactions are not required, as was the case with the previous two methods. The resulting products have the same 5′ end and a random 3′ end due to the incorporation of the fluorescently labelled ddNTP. The fluorescent dyes assign colours to the individual nucleotides (adenine – green, guanine – black, thymine – red, cytosine – purple).

The labelled DNA fragments are separated on a polyacrylamide gel. By passing the DNA fragment through the bottom of the gel, it is illuminated by a laser, whereby a fluorescent dye incorporated in the ddNTP emits light of a certain wavelength. A detector captures and analyses the wavelength of the emitted light and this information is then 'translated' into nucleotide sequence and processed by a computer (Figure 5.12).

DOI: https://doi.org/10.15414/2023.9788055226897

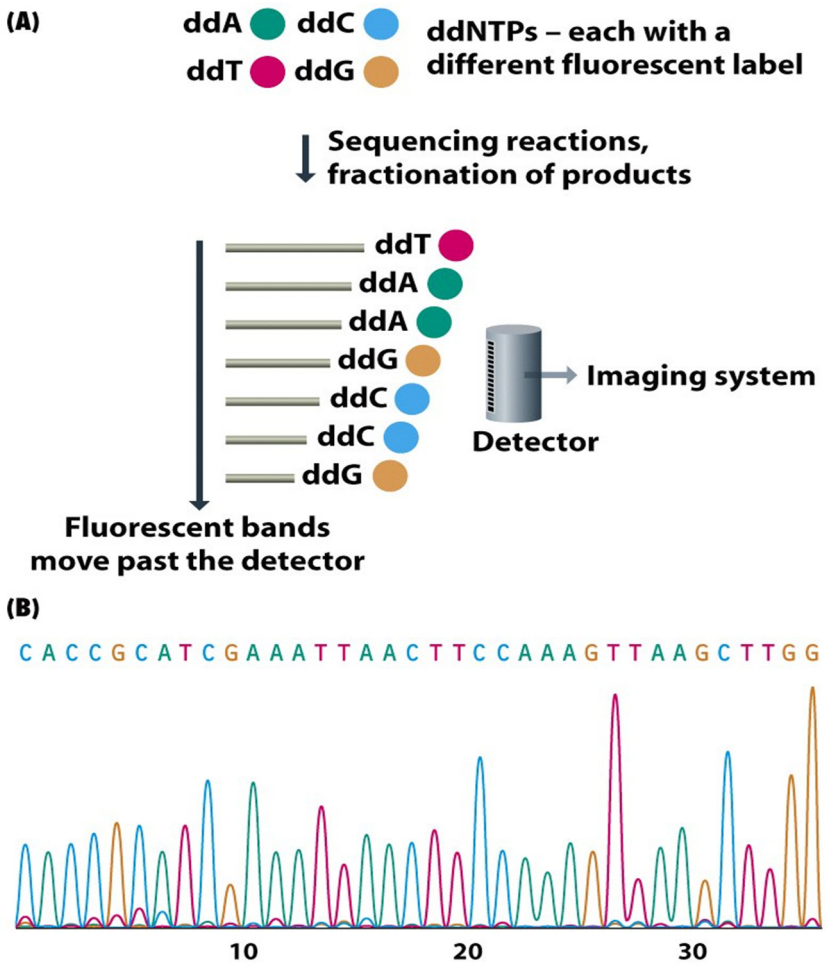


Figure 5.12 Reading the sequence generated by a chain termination experiment. (A) Each dideoxynucleotide is labeled with a different fluorophore. During electrophoresis, the labeled molecules move past a fluorescence detector, which identifies which dideoxynucleotide is present in each band. The information is passed to the imaging system. (B) A DNAsequencing printout. The sequence is represented by series of peaks, one for each nucleotide position. In this example, a green peak is an „A“, blue is „C“, brown is „G“, and red is „T“. (Brown, 2007, modified).

DOI: https://doi.org/10.15414/2023.9788055226897

5.3 Bioinformatics

Bioinformatics is a field focused on the development and application of computational, mathematical and statistical methods in the analysis of biological, biochemical and biophysical data. Genome and protein sequences are large biological data sets in volume, which can only be processed by appropriate computer and software equipment. Bioinformatics approaches are used to store biological data in a transparent way, to retrieve the necessary information and to analyse it afterwards.

Theoretical knowledge and practical experience in the field of bioinformatics is becoming an essential part of science and research. The established trend of progress in the field, presented by the development of specialized bioinformatics databases, new tools for in silico analyses, inevitably determines the search for further possibilities of archiving, cataloguing, data management and their evaluation. This trend is undoubtedly conditioned by the increasing level of technical and instrumental infrastructure, which at the same time brings with it an unfavourable phenomenon in the form of limited data quality and the amount of repetitive data.

Genomics and bioinformatics are currently one of the most intensively developing scientific fields. With their interdisciplinary content and application, they significantly influence several disciplines (genetics, biology, biochemistry, computer science, mathematics, medicine, pharmacology, agriculture and food science). The methods of genomics and bioinformatics make it possible to analyse the genomes and molecular-phenotypic interactions of organisms from a previously unrecognised perspective.

The source of bioinformatics is the complex analysis of genomes of non-cellular biological systems and organisms, in which methods of molecular biology are applied, especially the determination of the primary structure (sequencing) of nucleic acids and proteins. Genome and protein sequences are large biological data sets in volume, which can only be processed by appropriate computer and software equipment. Bioinformatics approaches are used to store biological data clearly, retrieve the necessary information and analyse it subsequently.

The main areas of bioinformatics include:

- information retrieval in databases,
- comparison of nucleic acid and protein sequences,
- discovery of gene functions in the genome (functional genomics),
- protein classification and proteomics,
- phylogenetic studies,
- comparison of genomes (comparative genomics).

In the field of agriculture and food, the application of bioinformatics results in several outputs:

- Identification of genes underlying production- and nutritionally valuable traits (genomics),
- studying the adaptability of the genome to abiotic and biotic stress factors (transcriptomics, proteomics, metabolomics),
- identification of factors determining the efficient use of soil nutrients (metagenomics),
- monitoring interactions between animal or human nutrition and gene expression (nutrigenomics).

The increase in the number and quality of biological data records has led to the development of a database management system. A database management system is the hardware and software that allows to provide all the required features of a database system and to work with the data.

Databases are an ordered set of information stored on a storage medium, including software that is designed to process and access the stored data. A number of organisations

around the world are involved in the collection and management of biological data, the development of tools for its analysis and the provision of information. On their websites, databases are publicly available that contain large datasets of nucleotide or amino acid sequences and other molecular biological properties. The information in the databases is continuously updated, sorted and accessible for searching and comparing nucleotide and amino acid sequences, genes, genomes, proteomes and phylogenetic studies.

5.3.1 Biological databases

Biological databases are biological data collected, managed, analysed and made publicly available on websites by specialised organisations. Biological databases include computer programs for updating and retrieving data. Biological databases contain different types of data, sequence, structural molecular, gene and protein expression, molecular interactions, mutational and phenotypic.

The data in the databases can be sequence, structural, molecular, gene and protein expression, molecular interactions, mutations and phenotypic variation, and others as listed in the general overview of the databases (Table 5.3).

In general, biological databases are oriented towards the following areas:

- DNA, RNA, protein sequences,
- domains, protein families,
- multidimensional structures of proteins,
- metabolic pathways,
- bibliographic data (publications),
- others (individual organisms and others).

Table 5.3 Division of databases by information content.

Database	Data type
Primary	Data obtained experimentally about the sequence of nucleic acids or proteins.
Secondary	Information derived from primary databases for genomes and metabolic pathways.
Metadatabases	They aggregate data from different sources and usually modify it into a more acceptable form, or focus on specific areas (data associated with specific manifestations of an organism).

Primary (nucleotide) databases ensure the availability of data and provide a means to retrieve additional information, such as the presence of similar sequences, the relationship to and between genes, the number of genetic data available for a given organism, and other information that arises from knowledge of the nucleotide sequence. They receive drafts of the original nucleotide sequence data, keep track of changes, and update the data on a daily basis to optimize synchronization of activity between databases.

The repository for the nucleotide sequence data of all organisms are three large databases. As the primary databases of original sequence data, they accept nucleotide sequence offerings, adding and exchanging new entries daily with the intention of optimally synchronizing their activities. Geographically, these databases are located in Japan, Europe and the USA.

The trio of primary DNA sequences consists of:

- DDBJ, the DNA Data Bank of Japan maintained by the National Institute of Genetics, Japan.
- EMBL, European Molecular Biology Laboratories, administered by the European Bioinformatics Institute, Hinxton, UK
- GenBank, the Gene Bank, managed by the National Center for Biotechnology Information (NCBI) in the USA.

The European Molecular Biology Laboratories (EMBL) and Nucleotide Sequence Database (NSD, also known as EMBL-Bank) is a database of primary nucleotide sequences. Together with the Sequence Read Archive (SRA) and the Trace Archive, it forms part of the European Nucleotide Archive (ENA). The European Bioinformatics Institute (EMBL-European Bioinformatics Institute, EMBL-EBI) is based in the United Kingdom. EMBL-Bank contains European primary nucleotide resources. The DNA and RNA sequence databases are directly oriented for individual research groups, genomic sequence projects and patent applications.

GenBank is a public database of nucleotide sequences on the American continent, also providing concise bibliographic and biological annotations. It is established and maintained by the National Center for Biotechnology Information (NCBI), part of the National Institutes of Health (NIH), based in Bethesda, USA.

DDBJ was created and is supported by the Center for Information Biology (CIB), the National Institute of Genetics (NIG) and the Japanese Ministry of Education, Culture, Sports, Science and Technology. DDBJ is based in Mishima, Japan.

All three primary nucleotide sequence databases (EMBL-Bank, GenBank, DDBJ) are officially associated in the International Nucleotide Sequence Database Collaboration (INSDC). INSDC ensures their mutual cooperation, exchange of complete database libraries and information.

5.3.2 Secondary databases

Secondary, or derived, nucleotide databases provide selected data types taken from primary nucleotide databases. Compared to primary databases, derivative databases contain more comprehensive documentation, such as detailed annotations, bibliographic data, links to review publications, and other supplementary information.

5.3.3 Specialised databases

Some biological databases focus on specific areas, such as data types, structures and functions of molecules, individual species of organisms, humans, and others. The classification of a database into different groups is not strict, but only an auxiliary overview criterion. Specialized databases include those that can be considered as primary or integrated. For example, the RefSeq (Reference Sequence Database) reference sequence database is a generic, integrated database with no redundancy and well-annotated genomics, transcript and protein sequence references. It is a database with primary, derived, specialized and integrated data. The specialization of biological databases is quite extensive and multifaceted. Databases target unique biological processes or other, not only informational molecules.

Examples of specialized databases:

- Carbohydrate structure databases (Carbohydrate structure databases), EuroCarbDB – carbohydrate structure information database.

- Protein-protein interactions databases, BIND Biomolecular Interaction Network Database, BioGRID A General Repository for Interaction Datasets (Samuel Lunenfeld Research Institute).
- Signal transduction pathway databases.
- Cancer Cell Map – signal transduction pathways of carcinogenesis.
- Netpath – managed sources of human signal transduction pathways.
- NCI – database of natural pathway interactions.
- Reactome – navigational map of human biological pathways.
- Metabolic pathway databases, BioCyc – a network of biological databases, MANET Molecular Ancestry Network – a bioinformatics database that maps the evolutionary relationships of protein structure.
- Microarray databases, ArrayExpress (European Bioinformatics Institute), Gene Expression Omnibus (National Center for Biotechnology Information), GPX (Scottish Centre for Genomic Technology and Informatics), Stanford Microarray Database (SMD) (Stanford University).
- Exosomal databases, ExoCarta.
- Mathematical model databases, BioModels Database – mathematical models of biological processes.
- Primer databases for PCR/RT PCR based methods (PCR /real time PCR primer databases), PathoOligoDB – oligo qPCR database for pathogen detection. Databases of different focus, Antibody Central – database of antibody information, CGAP Cancer Genes National Cancer Institute,
- Taxonomic databases, Catalogue of Life source databases, Encyclopedia of Life.

5.3.4 Integrated database systems and biological portals

Integrated or pooled databases and biological portals improve the user experience in a non-consumable number of biological databases. They significantly help to increase the efficiency of data retrieval in biological databases.

The Sequence Retrieval System (SRS) developed at the European Bioinformatics Institute (EBI) homogenises data from more than 80 biological databases. It has applications in databases of sequences, metabolic pathways, gene regulatory elements, 3D protein structures, mappings, mutations and locus-specific mutations.

Entrez is a molecular biology database and retrieval system. It was developed at the National Center for Bioinformatics (NCBI). It is an entry point for exploration in different but integrated databases.

DBGET is an integrated search system of major biological databases that are classified into five categories, KEGG databases in DBGET, other DBGET databases, web-only databases, linking web-only databases, and Pubmed databases.

RefSeq (Reference Sequence Database), a reference sequence database, is a general, integrated, non-redundant, well-annotated database of genomics, transcript, and protein sequence references.

OMIN - Online Mendelian Inheritance in Man (OMIN). PubMed - citations for biomedical literature. A freely available biomedical literature resource created and maintained by the National Center for Biotechnology Information (NCBI) and the U.S. National Library of Medicine (NLM) with localization at the National Institutes of Health (NIH).

Biology portals provide comprehensive information for working with biological databases, similar to an integrated biological database.

5.3.5 Bioinformatics tools

Comparing two or more sequences and finding the degree to which they are similar to each other is a central topic in practical bioinformatics. The use of digital symbols allows the similarity of two sequences to be not only qualitatively detected but also quantitatively measured. In many cases, to understand the relationships, the researcher must extend the analysis towards evolutionary relationships. Understanding phylogenetic relationships is essential for identifying genes, detecting the origins of genetic diseases, characterizing polymorphisms, and more. One of the key developments in bioinformatics has been the development of a procedure that allows sequences to be compared to determine their relatedness, named BLAST (Basic Local Alignment Search Tool) (Figure 5.13). That algorithm was developed by a collective of bioinformaticians led by Stephen F. Altschul in 1991. This type of comparative analysis allows the generation of interferences, for example, on the functional similarity between two proteins or on the content of similar motifs in their structure. BLAST is a local alignment algorithm that allows not only to record the region of best local alignment between the sequences being aligned (query) and the sequences or database (target sequences/target database) to which they are aligned, but also to identify other plausible alignments between the sequences being aligned. The wide applicability of this algorithm is based on the ability to accurately and quickly identify similarities between nucleotide or amino acid sequences. The aim of sequence comparison is to:

- find out their similarity,
- observe and record stable (conserved) and variable parts of the sequences,
- infer and evaluate evolutionary relationships between sequences.

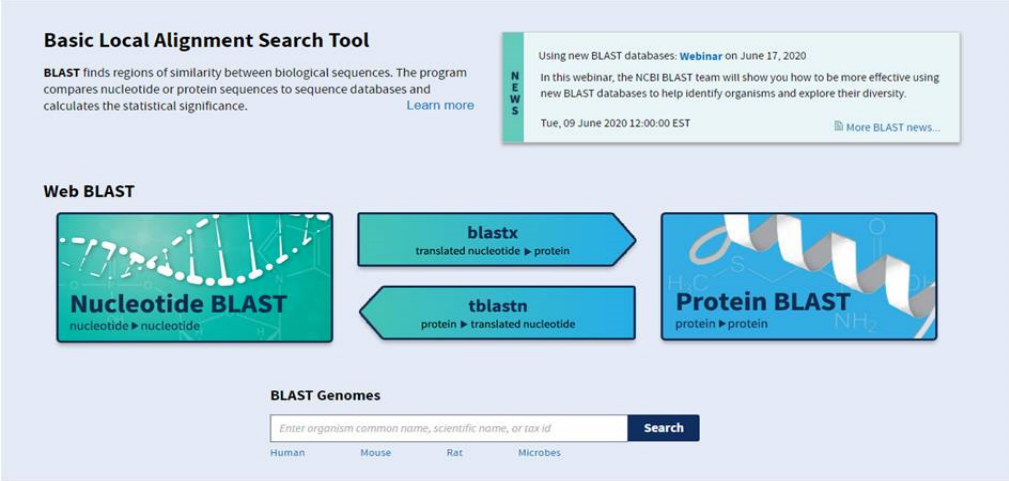


Figure 5.13 Home page of the BLAST algorithm of the National Center for Biotechnology Information (NCBI, https://blast.ncbi.nlm.nih.gov/Blast.cgi).

Several options of the above tool are available to the user, each implementing a specific type of sequence alignment.

A multiple sequence alignment is an alignment that contains more than two sequences. Multiple alignments increase the accuracy of this procedure between pairs of sequences, as well as patterns of conserved sequences (nucleotides or amino acids) that are not as obvious when comparing only two sequences. Multiple alignments are useful in defining protein structure and function and indispensable for phylogenetic analyses.

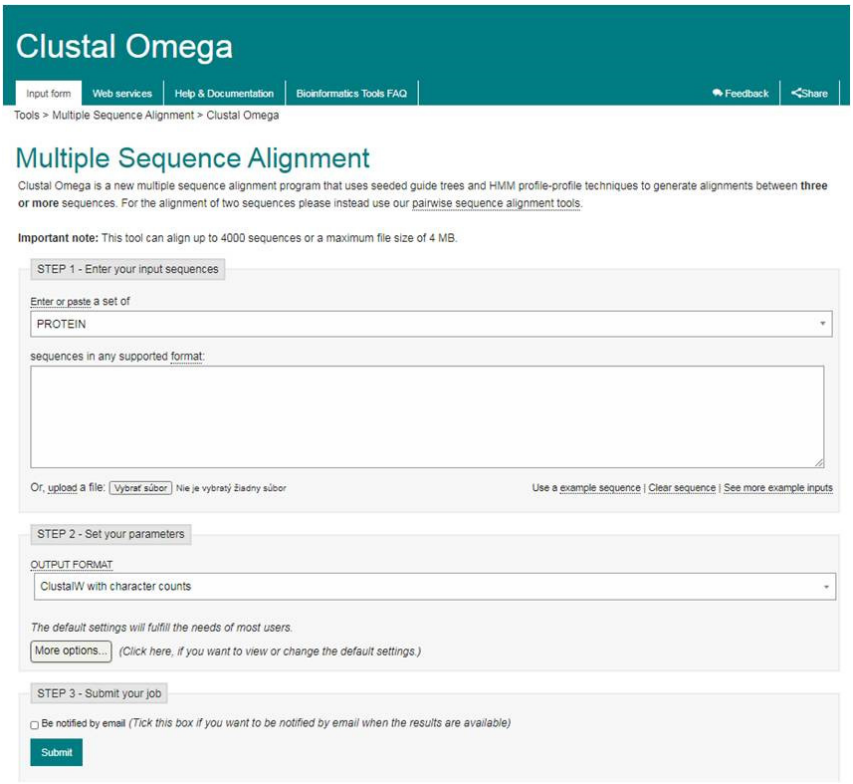


Figure 5.14 Homepage of the multiple sequence alignment tool, Clustal Omega. https://www.ebi.ac.uk/Tools/msa/clustalo/

Clustal Omega is the latest tool from the authors of the ClustalW hierarchical alignment software. It aligns three or more sequences together in a computationally efficient and accurate manner. It produces biologically meaningful multiple sequence alignments of divergent sequences. Evolutionary relationships are displayed through the Cladograms or Phylograms view. The Clustal Omega multiple sequence alignment tool is available at http://www.ebi.ac.uk/Tools/msa/clustalo/ (Figure 5.14).

Multiple sequence alignment highlights regions of similarity that may be related to specific features that were more conserved than other regions. These regions may in turn help to classify sequences or inform experimental design.

Aligning multiple sequences is also an important step for phylogenetic analysis, which aims to model substitutions that have occurred over evolution and to infer evolutionary relationships between sequences. The central idea of multiple alignment is the inclusion of amino acids and nucleotides into the same column because they are similar in some respect. The main criteria for creating a multiple alignment are given in Table 5.4.

DOI: https://doi.org/10.15414/2023.9788055226897

Table 5.4 Main criteria for creating multiple alignments.

Criteria	The meaning of the criterion
Structural similarity	Amino acids with the same function in each structure are placed in one column.
Evolutionary similarity	Amino acids or nucleotides related to the same amino acid (or nucleotide) in the common ancestor of all sequences are placed in a single column. Applied programs do not explicitly respect this criterion, but attempt to produce an alignment respecting this condition.
Functional similarity	Amino acids or nucleotides with the same function are placed in one column. Applied programs do not explicitly respect this criterion, but if the information is available, it is possible in some programs to select this option or to adjust the alignment manually.
Sequence similarity	Amino acids placed in a single column are those that achieved alignment with maximum similarity. Most programs use sequence similarity because it is the simplest criterion. If sequences are closely related, their structural, evolutionary, and functional similarity is equivalent to sequence similarity.

The first three criteria have obvious biological relevance. If the sequences are sufficiently similar, it is possible to use sequence similarity to create a multiple alignment that reflects the evolutionary, structural and functional relationships that exist between sequences.

DOI: https://doi.org/10.15414/2023.9788055226897

6 PROTEOMICS AND METABOLOMICS

6.1 Proteomics

Proteome is a set of all proteins of a certain system, such as cell organelles, cell, tissue, organ or organism. The proteome includes all functionally different forms of proteins. Functionally different forms are created by post-translational modification of a protein encoded by a single gene.

The Human Proteome Organization formulated the goals of proteomics in 2001 as:

a) identification of all proteins encoded by the human genome (or genomes of model organisms) with subsequent determination of their expression in various cells of the given organism - **expression proteomics**,
b) their subcellular localization in various organelles; their post-translational modification; their mutual interactions - **structural proteomics**
c) the relationship between structure and function – **functional proteomics**.

The cell proteome is influencing by many factors as it is showed in the figure 6.1.

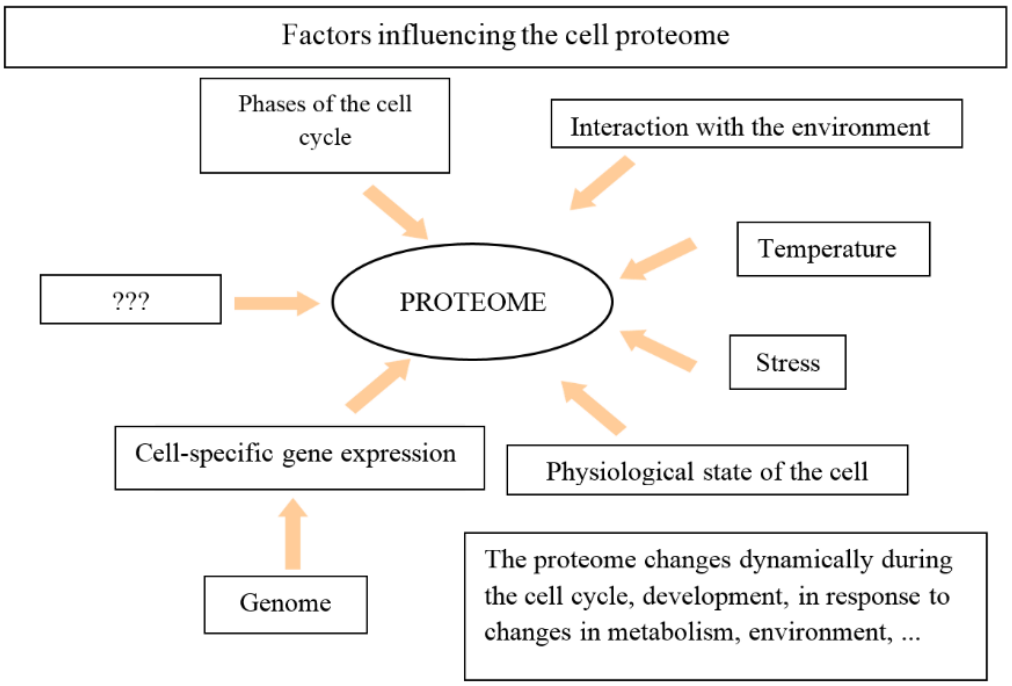


Figure 6.1 Factors influencing the cell proteome

6.1.1 Proteomics methods

1. Two-dimensional protein electrophoresis

Currently, two-dimensional protein electrophoresis is the method of choice for protein separation (Figure 6.2). This method enables the resolution of up to 10,000 proteins.

DOI: https://doi.org/10.15414/2023.9788055226897

The essence of the method is the use of two different physical and chemical properties of the protein: first, the proteins are divided according to their isoelectric point (pI), which is the pH value at which the sum of the charges of the protein is zero. This can be achieved by applying protein to the strip of gel in which the pH gradient is created and applying electrical voltage. The proteins then migrate to the cathode or anode according to their overall charge until the pH of the spot in the gel corresponds to the pI of the protein. After separation in the first dimension, ordinary electrophoresis on a polyacrylamide gel (PAGE) is performed, when the voltage is applied perpendicular to the original orientation of the electrodes. Proteins then migrate depending on their size. After both phases of 2D electrophoresis, the proteins need to be visualized chemically or radioactively. The resulting protein "maps" can be compared, e.g. between the experimental and control samples and thus identify the expressed proteins. Therefore, it is necessary to verify the identity of these specifically expressed proteins, most often by "cutting out" the area of the gel showing the difference and its subsequent analysis by means of mass spectrometry.

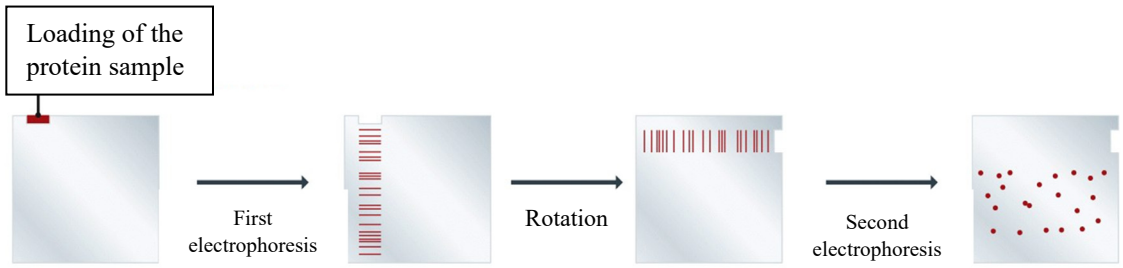


Figure 6.2 Two-dimensional protein electrophoresis

2. Mass spectrometry

Mass spectrometry is a method that allows accurate measurement of the molecular weight of a wide range of substances. Since the investigated substance must be transferred to the gas phase, the use of mass spectrometry for the analysis of proteins (but also polysaccharides and oligonucleotides) was made possible by the development of "soft" mass spectrometry ionization techniques, which include matrix-assisted desorption/ionization (MALDI) and electrospray ionization (ESI). Protein identification is carried out in two basic ways:

a) The protein is digested by trypsin or another proteolytic enzyme into smaller peptides, the exact masses of which are measured using mass spectrometry. The spectrum of these masses is then compared with the theoretical spectra that are calculated from the protein sequences in the available databases (using bioinformatics methods).

b) Tandem mass spectrometry makes it possible to select a peptide that is subsequently fragmented with an inert gas. The profile of the fragmentation result (fragmentation pattern) provides partial or complete information about the protein sequence, which is the basis for searching for a match with the data stored in the databases.

Sample preparation begins by mixing the sample with excess matrix (weak organic acid). The matrix will serve to absorb laser radiation. After being applied to the plate, the mixture is dried and inserted into the evacuated MALDI-TOF instrument. By irradiating the crystals of the mixture with a laser pulse, the sample is desorbed and ionized. The matrix acts as an ionizing agent, it evaporates and the volatile molecules of the sample pass into the gas phase. Subsequently, the movement of ions is accelerated by an electric field and they fly through the vacuum through a fly-by mass analyzer. The MALDI-TOF principle is based on

DOI: https://doi.org/10.15414/2023.9788055226897

the movement of a group of ions with different mass and charge ratios, placed in an electric field, which acquire the same kinetic energy (Figure 6.3). Their time of flight depends on the ratio of their mass and charge. The time of flight of ions is measured after the supply of kinetic energy in a region with zero electric field.

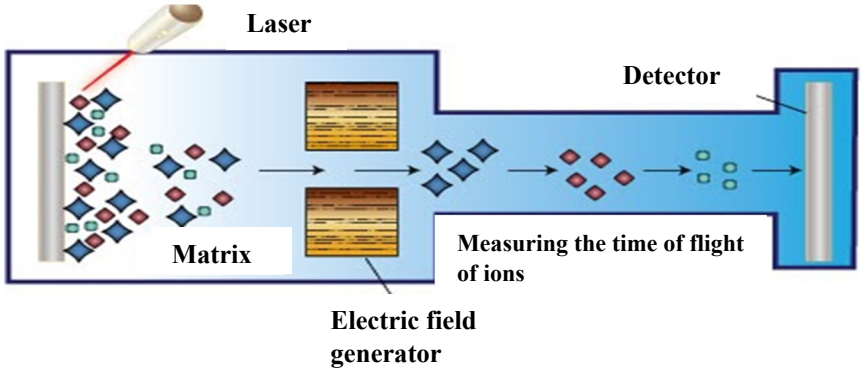


Figure 6.3 Principle of mass spectrometry by the MALDI-TOF method

3. Protein chip

Compared to nucleic acids, the variability of physicochemical properties of proteins is more complex and therefore the construction of a single "protein chip" (protein array) is far more complicated. Even so, there are currently several platforms available. Some use antibodies or antigens that are applied to the chip in high density and detect the corresponding entity in the sample, others contain non-protein polymer molecules that react with certain groups of proteins.

6.2 Metabolomics

The metabolome is a set of low-molecular substances of the cell, such as glucose, cholesterol, nucleotides, vitamins, substrates and intermediates of enzymatic reactions. Metabolomics studies low molecular weight substances in biological samples under various conditions, with the aim of determining similarities, differences, interactions, changes between components.

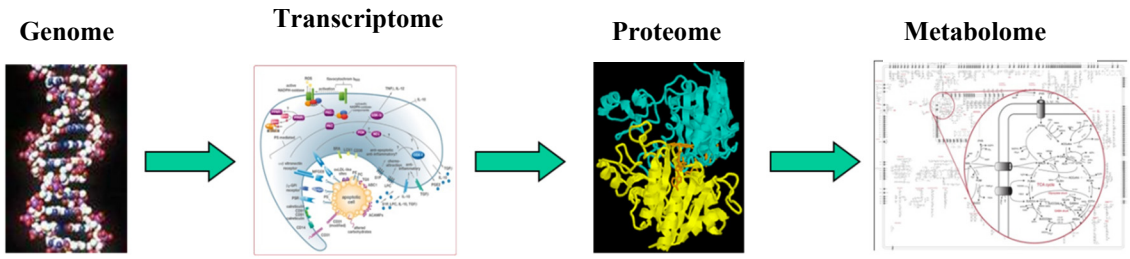


Figure 6.4 Interconnection of genome, proteome and metabolome.

Biochemical profiling of a cell or tissue in certain specific conditions provides an accurate characterization of biochemical processes that take place in different physiological

DOI: https://doi.org/10.15414/2023.9788055226897

phases of the organism, including pathogenic states (Figure 6.4). By converting the biochemical processes of a cell into a detailed set of metabolites, metabolomics provides a database of data that we can directly link with precise information coming from proteomic and genomic analyses. The metabolome can be characterized by chemical techniques, such as infrared spectroscopy, mass spectrometry, magnetic resonance spectroscopy, by means of which various low-molecular substances forming the products of the cell's metabolism are identified and quantified. The obtained information enables a consistent characterization of various biochemical processes, based on which it is possible to create models of the metabolic flows of individual metabolites of the cell. Subsequently, changes in the metabolome can be defined in terms of changes in the flow of metabolites, which provides a very sophisticated approach to identifying biochemically conditioned changes in a specific physiological stage of a cell, tissue or organism. Subsequently, the acquired knowledge can be used for metabolic engineering, where induced changes in the genome by mutation or recombinant DNA techniques lead to changes in biochemical processes in the cell in the desired direction, for example, an increase in the synthesis of antibiotics by microorganisms.

In order to systematically identify and quantify metabolites from a biological sample and achieve comprehensive characterization of biomarker targets, the analysis considers both endometabolome and exometabolome. Analysis of metabolomes can use two strategies: **untargeted** and **targeted** based on the objective of the study.

Untargeted metabolomics generates hypothesis and allows for full scanning of the metabolome and pattern identification.

Targeted metabolomics is generally performed for validation of an untargeted analysis. In the targeted approaches MS methods using standards are used for quantitative analysis.

Mass spectrometry is the mostly used analytical tool for identification and characterization of different biomolecules in biological sciences. There are various types of instrumentation and ionization methods which can be combined in order to obtain precise results. The standard MS method consists of basic steps:

1. Sampling – acquisition and preparation of samples
2. Separation – various chromatographic methods (LC, GC) are used to separate biological mixtures and sample can be directly injected to mass spectrometer
3. Ionization – based on samples and analytical method, the proper ionization source had to be selected: Electron ionization, chemical ionization, electrospray ionization, matrix assisted laser desorption ionization etc.
4. Detection – different mass spectrums based on mass to charge (m/z) are detected using various detectors: quadrupole, ion trap, time of flight, orbitrap.
5. Data analysis – raw analytical data had to be correctly interpreted

Targeted metabolomics used also nuclear magnetic resonance (NMR) spectroscopy to determine molecular structures and for quality control. NMR is the only detection technique which does not rely on separation of the analytes, and the sample can thus be recovered for further analyses. All kinds of small molecule metabolites can be measured simultaneously. The main advantages of NMR are high analytical reproducibility and simplicity of sample preparation. NMR can quantitatively analyze mixtures containing known compounds. Once the basic structure is known NMR is used to determine molecular conformation.

6.2.1 Applications of metabolomics

The metabolomics can be applied in many areas:

Functional genomics – predicting the function of unknown genes by comparison with the metabolic perturbations caused by deletion/insertion of known genes,

Agronomy – analyzing responses of plants to different stress factors can be used for efficient fertilization management

Nutrition – metabolic fingerprint, which reflects the balance of different factors (age, sex, genetics, diet, drugs etc.) on an individual's metabolism

Environmental metabolomics – studying interactions between organisms and environment

Toxicology – metabolic profiling can be used to determine physiological changes caused by chemicals and for disease diagnosis.

DOI: https://doi.org/10.15414/2023.9788055226897

6 REFERENCES

BARNHART, R.B.: Hammond Barnhart Dictionary of Science. Maplewood. New Jersey, Hammond 1986. 740 p.

BAXEVANIS, A.D., OUELLETTE, B.F.F.: Bioinformatics. A practical guide to the analysis of genes and proteins. 3rd ed., Hoboken : John Wiley & Sons, Inc., 2005, 540 p., ISBN 978-0-471-47878-2.

BAXEVANIS, A.D.: Searching the NCBI databases using Entrez. Current Protocols in Bioinformatics. Supplement 13. Hoboken : John Wiley & Sons, Inc. 2006.

BECKMANN, J. S., SOLLER, M. Toward a unified approach to genetic mapping of eukaryotes based on sequence tagged microsatellite sites. In: *BioTechnology*, vol. 8, 1990, p. 930-932.

BEŽO, M., BEŽOVÁ, K.: Genetický slovník. Nitra : SPU v Nitre, 1998, 318 s. ISBN 80-7137-556-X.

BEŽO, M., ŠTEFÚNOVÁ, V., ŽIAROVSKÁ, J., RAŽNÁ, K.: Genetika. Výkladový slovník. 1. vyd., Nitra : VES SPU v Nitre, 2013, 212 s. ISBN 978-80-552-1072-8

BEŽO, M., ŽIAROVSKÁ, J., RAŽNÁ, K., ŠTEFÚNOVÁ, V.: Metódy genetických technológií. Praktikum. 1. vyd., Nitra : VES SPU v Nitre, 2015, 111s. ISBN 978-80-552-1293-7.

BLAST Basic Local Alignment Search Tool. [online]. Dostupné na internete: http://blast.ncbi.nlm.nih.gov/.

BLAST® Help manual. https://www.ncbi.nlm.nih.gov/books/NBK1762/.

BREJOVÁ, B., VINAŘ, T.: Metódy v bioinformatike. 1. vyd., Bratislava : Knižničné a edičné centrum Univerzita Komenského, 2011, 92 s. ISBN 978-80-89186-94-5

BROWN, T.A. Genomes 3. New York: Garland Science Puglishing, 2007. 713 p. ISBN 0-8153-4138

CLAVERIE, J.M., NOTREDAME, C.: Bioinformatics for dummies. 2nd ed., Hoboken : Wiley Publishing, Inc., 2007, 436 p. ISBN 978-0-470-08985-9.

COLLARD, B.C. & MACKILL, D.J. Start Codon Targeted (SCoT) Polymorphism: A Simple, Novel DNA Marker Technique for Generating Gene-Targeted Markers in Plants. In: *Plant Molecular Biology Reporter,* vol. 27, 2009*,* no. 1, p. 86-93.

COOKE, R. J. Allelic variability at the Glu-1 loci in wheat varieties. In: *Plant Varieties and Seeds*, vol. 8, 1995, p. 97 – 106.

CVRČKOVÁ, F. Úvod do praktické bioinformatiky. 1. vyd., Praha : Academia, 2006, 148 s. ISBN 80-200-1360-1.

ČERNÝ, J., ŠASEK, A. Analýza genetickej štruktúry krajových odrôd pšenice pomocou signálnych gliadínových gliadínových a glutenínových génov. In: *Scientia Agriculture bohemica*, roč. 27, 1996, č. 3, s. 161 – 182.

ČERNÝ, J., ŠASEK, A. Elektroforetická analýza gliadínov s vysokou molekulovou hmostnosťou odrôd pšenice (*Triticum aestivum* L.) testovaných v štátnych odrodových pokusoch v Rakúsku. In: *Scientia Agriculture bohemica*, roč. 27, 1996, č. 4, s. 237 – 260.

ČERNÝ, J., ŠASEK, A., MALÝ, J. Ověření metody bílkovinných markerú pekařské jakosti pšenice obecné pomocí nových genotypú, zkoušených ve státních odrúdových zkouškách v roce 1991. In: *Genetika a Šľechtení*, roč. 28, 1992, č. 4, s. 271 – 283.

ČERNÝ, J., ŠAŠEK, A., VEJL, P., HANIŠOVÁ, A. Common wheat (*T. aestivum* L.) marking by determination of approximate dependence of frequency of allelic gliadin genes on quality grade of agronomic character. In: *Scientia Agriculturae Bohemica*, vol. 26, 1995, no. 4, p. 245 – 258.

DELLAPORTA, S. L., WOOD, J., HICKS, J. B. A plant DNA minipreparation: Version II. In: *Plant Mol. Rep*., vol. 4, 1993, p. 19 – 21.

DRAPER, S. R. ISTA variety committee. Report of the working group for biochemical tests for cultivar identification 1983-1986. In: *Seed Sci. Technol*., 1987, p. 431 – 434.

GÁLOVÁ, Z., BALÁŽOVÁ, Ž., CHŇAPEK, M., VIVODÍK, M., OSLOVIČOVÁ, V.: Bielkovinové a DNA markery pšenice. 1. vyd. Nitra: Slovenská poľnohospodárska univerzita v Nitre, 2011, 175 s. ISBN : 978-80-552-0673-8.

GÁLOVÁ, Z., PALENČÁROVÁ, E., CHŇAPEK, M., BALÁŽOVÁ, Ž.: Využitie obilnín, pseudoobilnín a strukovín v bezlepkovej diéte. 1. vyd. Nitra : Slovenská poľnohospodárska univerzita, 2012, 172 s. ISBN 978-80-552-0826-8

GÁLOVÁ, Z., SMOLKOVÁ, H., MICHALÍK, I. HMW – glutenínové podjednotky ako markery chlebopekárskej kvality zrna pšenice. In: Využitie integrovanej rastlinnej výroby v podmienkach Slovenska. Nitra: Vysoká škola poľnohospodárska, 1996, s. 248 – 250.

GÁLOVÁ, Z., MICHALÍK, I., KNOBLOCHOVÁ, H., GREGOVÁ, E. Variation in HMW glutenin subunits of different species of wheat. In: *Rostlinná výroba*, vol. 44, 2002, p. 111-116.

GENOVA: BEŽO, M., RAŽNÁ, K., BEŽOVÁ, K., ŠTEFÚNOVÁ, V., KUTIŠOVÁ, J., ŽIAROVSKÁ, J. Genetické inžinierstvo rastlín v obrazoch [elektronický zdroj]. 1. vyd. Nitra : Slovenská poľnohospodárska univerzita, 2005. 1 CD-ROM (10,44 AH). ISBN 80-8069-627-6.

GOODACRE, R., VAIDYANATHAN, S., DUNN, W. B., HARRIGAN, G. G., KELL, D. B.. Metabolomics by numbers: acquiring and understanding global metabolite data. In: *TRENDS in Biotechnology*, vol. 22, 2004, p. 245-252.

GREGÁŇOVÁ, Želmíra: Molekulárne markery v identifikovani a charakteristike pšenice letnej : dizertačná práca. Nitra : Slovenská poľnohospodárska univerzita v Nitre, 2005, 142 s.

GREGOVÁ, E., KRAIC, J., ŽÁK, I. Charakterizácia odrôd pšenice pomocou glutenínov. In: Biochemické, molekulárne a morfologické techniky v identifikácii odrôd rastlín. Bratislava: ÚKSUP, 1995, s 11 – 14.

GUIRINEAU, F., LUCY, A., MULLINEAUX, P. Effect of two consensus sequences preceeding the translation iniciation codon on gene expression in plant protoplasts. In: *Plant Mol. Biol.*, vol. 18, 1992, p. 815-818.

HADFIELD, K.A., DANDEKAR, A. M., ROMANI, R.J.: Demethylation of ripening specific genes in tomato fruit. In: *Plant science limerick*, vol. 92, 1993, no. 1, p. 13-18.

HARTL, D. L., JONES, E. W.: Essential Genetics. 2nd ed. USA : Jones and Bartlett Publishers, 1999, 552 pp. ISBN 0-7637-0838-0

HOLIDAY, R.: The inheritance of epigenetic defects. Science (Washington, DC), 223, 1987, p. 163-170.

HOPSON, J. L., WESSELES, N. K.: Essentials of biology. New York : McGraw – Hill Publishing Company, 1990. 865 p. ISBN 0-07-557108-0

CHAMBERS, G.K., MACAVOY, E.S.. Microsatellites: consensus and controversy. In: *Comparative Biochemistry and Physiology Part B*, vol. 126, 2000, p. 455-476.

CHAVLA, H. S.: Introduction to plant biotechnology. Science Publishers, Inc., New Hampshire, USA, 2004, 538 p. ISBN 1-57808-228-5

CHRISTOU, P., KLEE, H.: Handbook of Plant Biotechnology. John Wiley&Sons, Ltd., Great Britain, 2004, 1420 p. ISBN 0-471-85199-X

JABLONKA, E. Inheritance systems and the evolution of new levels of individuality. In: *Journal of theoretical biology*, vol. 170, 1994, no. 3, p. 301-309

KOLSTER, P., VEREIJKEN, J. M. Evaluating HMW – glutenin subunits to improve breadmaking, quality of wheat. In: *Cereal fd Wld*, 1993, č. 2, s. 76 – 83.

KRAIC, J.: Molekulárna diferenciácia a charakterizácia genotypov rastlín : Doktorandská dizertačná práca. Piešťany : UKF, 1999. 120s.

KRAIC, J., FARAGÓ, J., OSTROLUCKÁ, M. G., LIBANTOVÁ, J., MORAVČÍKOVÁ, J., JOMOVÁ, K. HRAŠKA, Š.: Biotechnológie rastlín. Nitra: UKF, 2011, 320 s. ISBN 978-80-8094-885-6

KRAIC, J., GREGOVÁ, E., HERMUTH, J. Protein heterogeneity in European wheat landraces and obsolete cultivars. In: *Genetic Resources and Crop Evolution*, vol. 2, 1999, p. 1- 8.

LESK, A.M.: Introduction to genomics. 1st ed. New York : Oxford University Press Inc., 2007, 419 p. ISBN 978-0-19-929695-8.

MACAS, J., MÉSZÁROS, T., NOUZOVÁ, M. Plantsat: a specialized database for plant satelite repeats. In: *Bioinformatics*, vol. 18, 2002, no. 1, p. 28–35.

MANSON, A. L., JONES, S., MORRIS, A.: Cell Biology and Genetics. New York : Elsevier Science Ltd., 2002, 232 p. ISBN 0723432481.

MARHOLD, K., HINDÁK, F.: Zoznam nižších a vyšších rastlín Slovenska. Bratislava : Veda, 1998, 687 s. ISBN 80-224-0526-4

MASOJC, P. 2002. The application of molecular markers in the process of selection. In: *Cellular & Molecular Biology Letters*, vol. 7, 2002, p. 499-509.

MOHAMMADI, M.M., BAVI, O. DNA sequencing: an overview of solid-state and biological nanopore-based methods. In: *Biophys Rev.,* vol. 14, 2021, p. 99-110.

MOUNT, D.V.: Bioinformatics. Sequence and genome analysis. 1st ed., New York : Cold Spring Harbour Laboratory Press., 2001, 560p. ISBN 290-0-879-69608-4

MROZEK, D., MALYSIAK-MROZEK, B., SIĄŻNIK1, A. Search GenBank: interactive orchestration and ad-hoc choreography of Web services in the exploration of the biomedical resources of the National Center For Biotechnology Information. In: *BMC bioinformatics*, vol. 14, 2013, no. 73.

ONGA, Q., NGUYENC, P., THAOC, N.P., LEC, L. Bioinformatics Approach in Plant Genomic Research. In: *Current Genomics*, vol. 17, 2016, p. 368-378.

OSTELL, J.: The Entrez search and retrieval system. The NCBI Handbook. Chapter 15. 2003.

OXFORD DICTIONARY OF BIOLOGY.: 4th ed. Oxford : Oxford University Press Inc., 2000. 641 p. ISBN 0-19-280102-3.

PAYNE, P. I. 1987. Genetics of wheat storage proteins and the effect of allelic variation on bread-making quality. In: *Ann. Rev. Plant Physiol*, vol. 38, 1987, p. 141-153.

PAYNE, P. I., NIGHTINGALE, M. A., KRATTIGER, A. F. The relationship between the HMW glutenin subunit composition and the bread-making quality of British-grown wheat varieties. In: *Cereal Research Communications*, vol. 11, 1983, no. 6, p. 29–35.

POWELL, W., MORGANTE, M., ANDRE, C. et al. The comparison of RFLP, RAPD, AFLP and SSR (microsatellite) markers for germplasm analysis. In: *Molecular Breeding*, vol. 2, 1996, p. 225-238.

RÖDER, M. S., KORZUN, V., WENDEHAKE, K. et al. A microsatellite map of wheat. In: *Genetics*, vol. 149, 1998, p. 2007–2023.

SAIKI, R. K., SCHARF, S., FALOONA, F. et al. Enzymatic amplification of beta – globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. In: *Science*, vol. 230, 1985, p. 1350-1354.

SEMAGN, K., BJORSTAND, A., NDJIONDJOP, M.N. An overview of molecular marker methods for plants. In: *African Journal of Biotechnology*, vol. 5, 2006, p. 29.

SNUSTAD, D. P., SIMMONS, M.J.: Principles of genetics. 6th ed. John Wiley & Sons, Inc., Hoboken : USA, 2012. 767 p. ISBN 978-0-470-90359-9

SHEWRY, P. R., HALFORD, N. G., TATHAN, A. J. The HMW subunits wheat, barley and rye. In: *Oxford Suveys of Plant Molecular Cell Biology*, vol. 6, 1989, p. 163 – 219.

ŠAŠEK, A., ČERNÝ, J., BRADOVÁ, J., NECVETAJEV, V. P. A catalogue of electrophoretic hordein spectra in the assortment of winter barley varieties and new varieties. In: *Scientia Agric. Bohemoslov*, vol. 22, 1990a, p. 11 – 22.

ŠAŠEK, A., ČERNÝ, J., KRAUTOVÁ, E. Hordeínová spektra vybraných odrúd československého sortimentu ječmene jarního. In: *Genet. A Šlecht*, vol. 24, 1988, p. 47-56.

TAYLOR, J.L., JONES, J.D.G., SANDLER, S., MUELLER, G.M., BEDBROOK, J., DUNSMUIR, P. Optimizing the expression of chimeric genes in plant cells. In: *Mol. Gen. Genet.,* vol. 210, 1987, p. 572-577.

THAIM, M., HICKMAN, M.: The Penguin dictionary of biology. 9th ed. Harmondsworth, Middlesex: Penguin Books Ltd, 1996, 665 p. ISBN 0-14-051288-8

TODOROVSKA, E. Retrotransposons and their role in plant-genome evolution. In: *Biotechnol. & Biotechnol. Eq.*, vol. 21, 2007, p. 294–305.

TYERS, M, MANN, M. From genomics to proteomics. In: *Nature*, 2003, vol.13, p. 193-197.

VICIENT, C. M., JAASKELAINEN, M. J., KALENDAR, R., & SCHULMAN, A. H. Active retrotransposons are a common feature of grass genomes. In: *Plant Physiology*, vol. 125, 2001, p. 1283-1292.

VICIENT, C. M., SCHULMAN, A. H. Copia-Like Retrotransposons in the Rice Genome: Few and Assorted. In*: Genome Lett.*, vol. 1, 2002, p. 35–47.

VOET, D., VOET, J. G.: Biochemistry, John Wiley & Sons, Inc., 1990, 1323 p.

VOS, P., HOGERS, R., BLEEKER, M. et al. AFLP: a new technique for DNA fingerprinting. In: *Nucl Acid Res*, vol. 23, 1995, p. 4407-4414.

WALKER, J. M., RAPLEY, R.: Molecular biology and biotechnology. 4th ed., Cambridge : The Royal Society of Chemistry, 2000, 563 p. ISBN 0-85404-606-2

WEAVER, R. F., HEDRICK, P. W.: Genetics. 3rd ed., Dubuque : Wm. C. Brown Publishers, 1997, 638 p. ISBN 0-697-16000-9

WELSH, J., MCCLELLAND, M. Fingerprinting genomes using PCR with arbitrary primers. In: *Nucleic Acids Res*, vol. 18, 1990, p. 7213-7218.

WILLIAMS, J. G. K., KUBELIK, A. R., LIVAK, K. J. et al. DNA polymorphism amplified by arbitrary primers are useful as genetic markers. In: *Nucl Acid Res*, vol. 18, 1990, p. 6531-6535.

WRIGLEY, C. W. Identification of cereal varieties by gel electrophoresis of the grain proteins. In: Linskens, H. F., Jackson, J. F.: Seed Analysis., Berlin, Heilderberg, Springer-Verlag, 1992, p. 17 – 41.

http://www.ncbi.nlm.nih.gov/

http://www.ebi.ac.uk/services/

http://molbiol-tools.ca/

http://123genomics.homestead.com/files/analysis.html

http://www.hort.purdue.edu/hort/courses/HORT250/

http://europa.eu/scadplus/leg/en/s86000.htm

http://www.avatar.se/lectures/strbio2001/databases/nuc.html

https://www.nlm.nih.gov/pubs/techbull/so03/so03_global_query.html>

http://www.biodbs.info/Dh.html>

http://bioinformaticsweb.net/datalink.html

http://www.rbgkew.org.uk/cval/homepage.html

http://dna-barcoding.blogspot.com/2013/08/maldi-tof-ms.html

https://acsess.onlinelibrary.wiley.com/doi/full/10.3835/plantgenome2008.03.0181

Title: Biotechnology in Plant Production II

Authors: Ž. Balážová, Z. Gálová, M. Vivodík, M. Chňapek, K. Ražná, J. Libantová

Publisher: Slovak University of Agriculture in Nitra

Edition: first

Year of publication: 2023

Form of publication: online

Pages: 72

AQ – PQ: 4.67 – 5.82

Not reviewed at the Publishing House of SUA in Nitra.

ISBN 978-80-552-2689-7

DOI: https://doi.org/10.15414/2023.9788055226897