Identification of Species with DNA-Based Technology

- All methods for the identification of species that rely on DNA or protein sequence analysis presuppose the neutral theory of molecular evolution, in which different lineages diverge over evolutionary times by the accumulation of molecular changes (most of them neutral).

- These methods are based on the assumption that individuals from a same species carry specific DNA (or protein) sequences that are different from those found in individuals from other species.

- However, the distribution of a given molecular variant in time and in space will be influenced by the reproductive success of individuals, migratory events and random genetic drift.

- **It is also important to keep in mind that there is no perfect DNA-typing method and that the choice of a particular technique is often a compromise that depends on a number of factors.**

- A number of different techniques are available for identifying genetic differences between organisms. **The choice of technique for any one specific use will depend upon the material being studied and the nature of the questions being addressed.** Molecular techniques differ in the way they sample within the genome and in the type of data that they generate.

PRE-DNA WORLD" IN SPECIES IDENTIFICATION PROCEDURES

- The term "morphology" is used in biology to refer to the form and structure of an organism as a whole or its component parts.
- It is unquestionable that the use of external or internal features of an individual is still the most applied process in both identification and taxonomy.

- Occasionally, the identification procedure is supported by behavioural characteristics (difference in habitat preferences, breeding seasons, epidemiology, etc) or physiological features (growth rates, biochemical composition, etc).

- Although extremely useful in several cases to assign organisms to well-defined categories, the use **of anatomical characters for species identification procedures has several disadvantages.**

- **First, there is a considerable morphological plasticity between** organisms of the same species. For instance, **coloration** in some species of birds and fishes are known to vary due to different **nutritional regimes**. For that reason, a reliable diagnostic procedure can be time-consuming and require the expertise of different taxonomists (when available).

- The use of morphology is also complicated **by the existence of sibling species - species that are morphologically nearly identical but are nonetheless reproductively isolated from one another.**
- Groups of **closely related species** in this condition can form large **cryptic species complexes.**

- A recent report demonstrates that these complexes are almost evenly distributed among major **metazoan taxa**. Moreover, these methods are hampered by the existence of convergent evolution, in which the same phenotypic feature can emerge independently in phylogenetic unrelated organisms. This fact can led to erroneous identifications if a small number of morphological features are considered in the analysis.

- Finally, most morphology-based approaches cannot be applied in cases where there is just a small amount of biological material available for examination.

- It was only in the second half of last century that, with the convergence of new ideas from genetics and biochemistry and a set of new technological developments, the field of species identification started to rely on information from the molecular components of cell.

- The first molecular methods successfully employed **were based on the analysis of proteins: protein** sequencing, protein electrophoresis, isoenzyme analysis, immunological reactions.

- Although each method has its own advantages, **a number of features are known to limit the use of proteins:**

- its rapid degradation in samples under stress of cross-reactions with proteins from closely related species, the differential expression of proteins in specific tissues, the scarcity of available antibodies for immunological reactions.

THE "DNA REVOLUTION" IN MOLECULAR SPECIES IDENTIFICATION

- **Three major characteristic of the DNA molecule makes it an extremely useful tool for molecular species identification.**
- First, DNA is an extremely **stable and long-lived biological molecule that can be recovered from biological material that has been under stress conditions** (processed food products, coprolites, mummified plant tissues, blood stains, etc).

- **Second, DNA is found in all biological tissues or fluids with nucleated cells** (or non-nucleated cells with plastids and/or mitochondria), enabling its analysis from almost all kinds of biological substrates (saliva, faeces, plant seeds, milk, etc).

- **Finally, DNA can provide more information than proteins due to the degeneracy of the genetic code and the presence of large non-coding stretches.**

Restriction Fragment Length Polymorphisms (RFLPs).

- The RFLP analysis **is widely used for the detection of interspecies variation at the DNA sequence level.** It consists in the generation of species-specific band profiles through ***the digestion of DNA with one or more restriction endonucleases*****).**

- These **restriction enzymes** cleave the DNA molecule at specific **4-6 base pair (bp) recognition sites, originating a set of fragments with different lengths that could be separated according to their molecular size by conventional gel electrophoresis.**

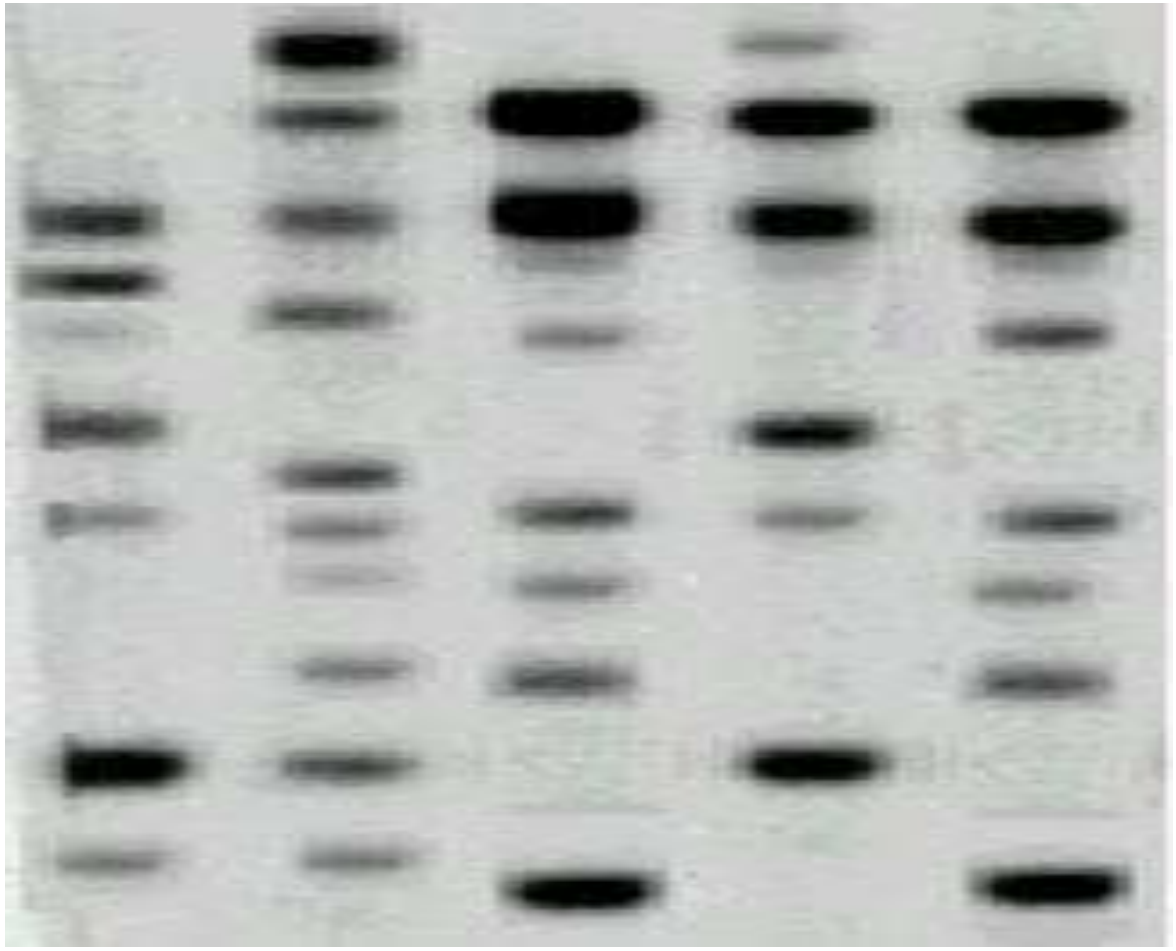
- The **RFLP banding pattern** could be visualized by hybridizing restriction fragments with a labelled probe in a solid support (for instance, by Southern blotting) or by treating the electrophoretic gel with ethidium bromide or silver staining.

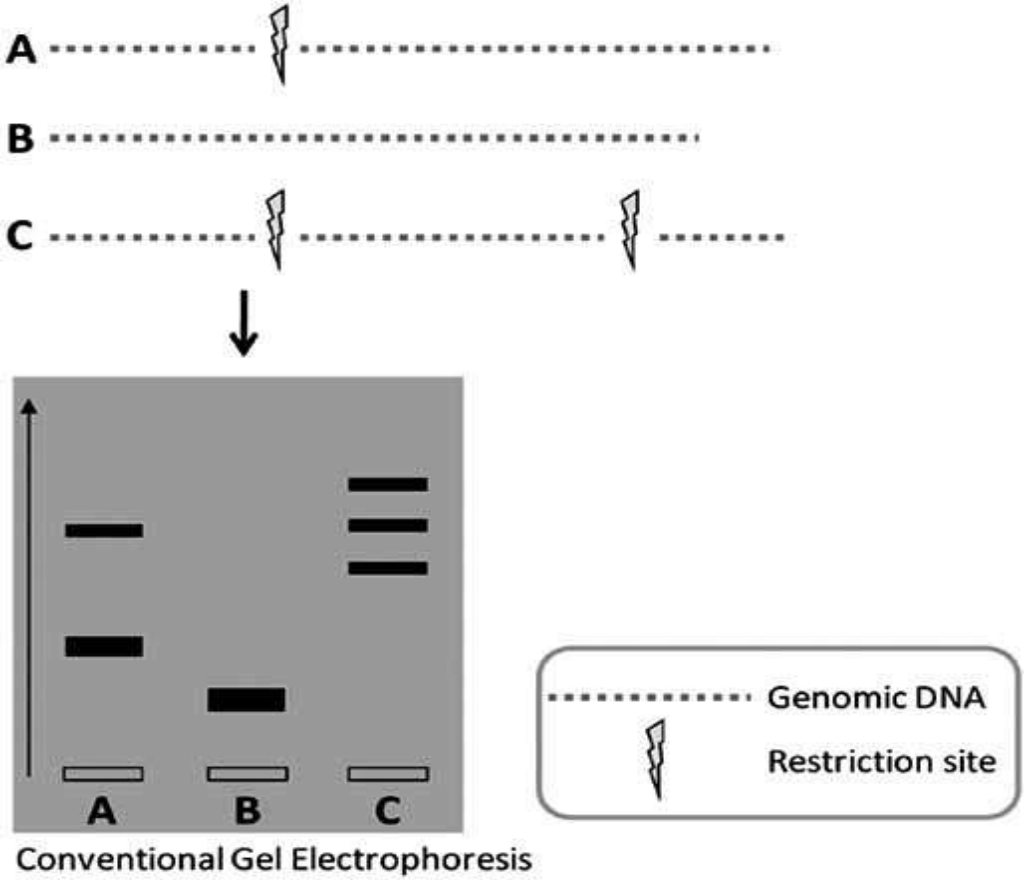
- **The distinctive RFLP profile** of each species is the result of the unique genomic distribution of recognition sites (generated or removed by single-base substitutions) and the distance between them (that varies due to large genomic rearrangements, such as translocations, transposable elements or tandem duplications).

- **A major disadvantage of the RFLP technique** is the possible existence of intraspecies mutations at restriction sites that can lead to false results due to the gain or loss of restriction fragments.

- This method relies on just a few informative DNA sequence positions, meaning **that several restriction enzymes are usually required to achieve a correct identification.** In those situations, the use of different enzymes generates highly complex RFLP patterns of difficult interpretation.

- Moreover, it is not amenable for automation and standardization because it requires a substantial amount of high quality DNA (unless a whole genome amplification has been performed prior to the RFLP analysis).





Schematic representation of the Restriction Fragment Length Polymorphisms (RFLPs) method. Genomic DNA is digested by restriction enzymes, originating a set of fragments with different lengths. Species (A, B and C) are identified by running the restriction fragments on a conventional electrophoretic gel.

notes

- **Mitochondrial DNA (mtDNA)**

- **The mitochondrial genome consists of a double-stranded DNA molecule devoted to the coding of key subunits of the electron transport chain found in mitochondria** (the powerhouses of eukaryotic cells). With few exceptions, all eukaryotic species have mitochondria. The mitochondrial genome of animals and plants are known to evolve at different rates.

- The typical animal mtDNA **has a high mutation rate and an exceptional organizational economy, with rare non-coding segments.** In contrast, mitochondrial genomes found in plants have large amounts of non-coding segments and a low accumulation of diversity.

- **The accelerated evolutionary rate of animal mtDNA** (and also of certain fungi and protists species) **implies that significant amounts of sequence variation could be found in closely related species – a useful feature for species identification procedures.** Moreover, in most species, mtDNA is uniparentally inherited without recombination, a fact that greatly simplifies the interpretation of results.

- The mtDNA is also easier to retrieve from low-quantity and/or degraded DNA samples since it is present in many copies per cell, providing a clear advantage over nuclear genome-based methods.
- The most important limitation of using mtDNA information **in the definition of species is the putative occurrence of male-biased gene flow between species (in cases where the mtDNA is maternally inherited).**

Polymerase chain reaction (PCR)

- The development of the polymerase chain reaction (PCR) technique has significantly improved the efficiency of laboratorial diagnostic procedures by allowing the *in vitro* formation of a large number of DNA copies (amplification) using a specific genomic region as template.
- Since it only requires a small amount of template DNA, the PCR method could be particularly useful for the identification of species in suboptimal DNA samples (processed food products, forensic samples, archaeological remains, etc).

Ribosomal RNA genes

- The genomic organization of ribosomal RNA (rRNA) genes **- responsible for the synthesis of RNA species** (the core of ribosomes) - is slightly different in prokaryotes and eukaryotes. Most **Bacteria and Archaea contain either a single or multiple copies of rDNA clusters dispersed in the genome.** Each cluster includes **the 16S, 23S and 5S rRNAs** separated by the internal transcribed spacer (ITS) regions and flanked by the 5′ and 3′ external transcribed spacers (5′-ETS and 3′-ETS).

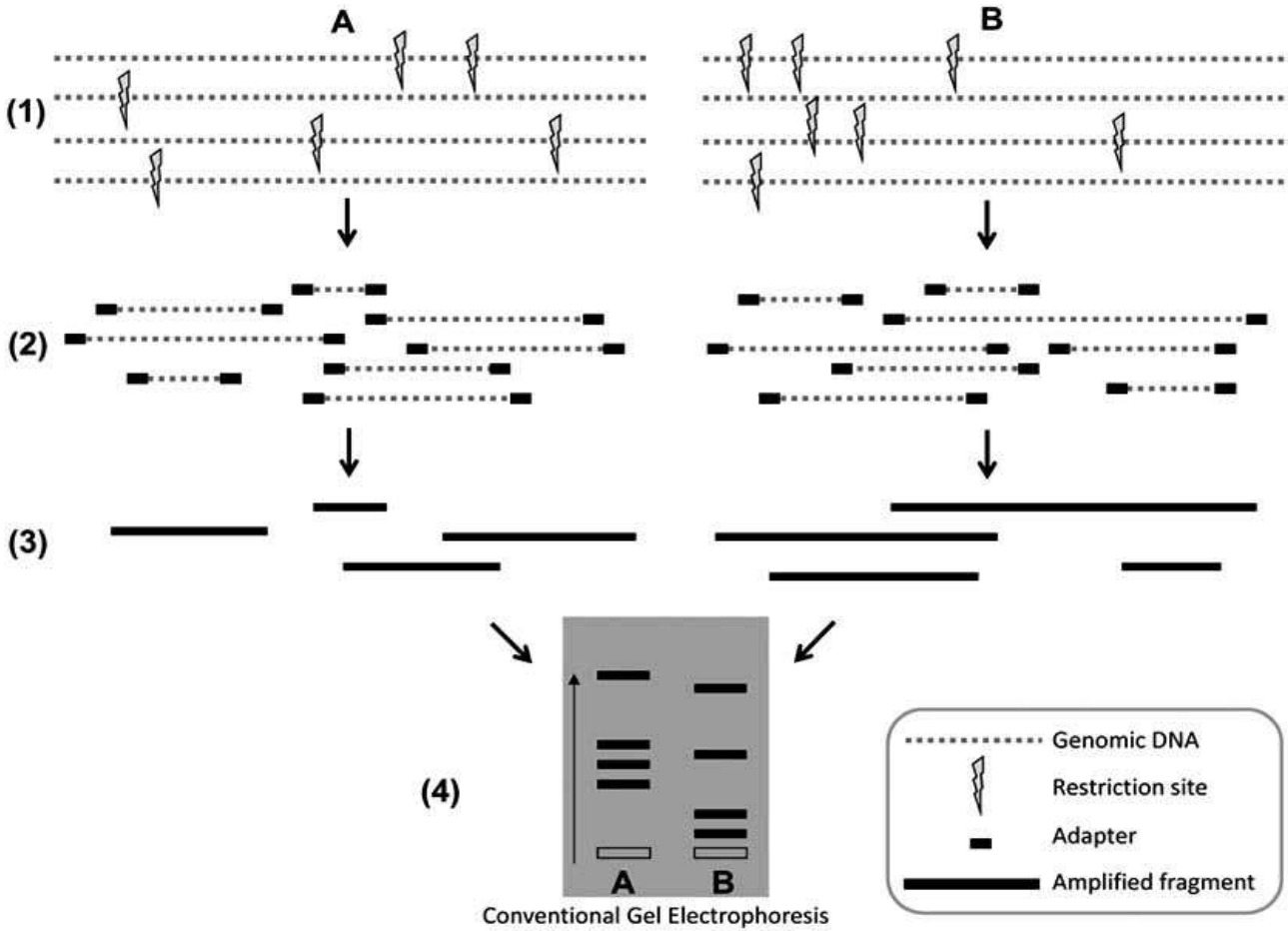
- Almost all eukaryotes have several copies of each rDNA cluster organized in tandem repeats. In this case, each cluster contains the 18S, 5.8S and 25/28S rRNAs, while the 5S gene is present in separate repeat arrays in the majority of eukaryotes.

Amplified Fragment Length Polymorphisms (AFLPs).

The AFLP method combines the reproducibility of restriction fragment analysis with the power of PCR. It is based on the selective PCR amplification of restriction fragments from a total digest of genomic DNA .The method usually works by digesting a small amount of purified genomic DNA **with two or more restriction enzymes (such as *Eco*RI and *Mse*I).**

- Double-stranded oligonucleotide adapters (10-30 bp long) are ligated to the sticky ends of DNA fragments (both 5′ and 3′ ends) generated during the restriction digestion.
- The ligated DNA fragments are then amplified twice under highly stringent conditions by PCR using primers complementary to the adapter and restriction site sequence.
- These selective primers include additional nucleotides at their 3′ end to reduce the complexity of the mixture of fragments. For instance, a selective primer with

- **The AFLP technique permits** the simultaneous screening of different loci randomly distributed throughout the genome. However, it is technically demanding in the laboratory, labour consuming and the interpretation of results may need automated computer analysis. Additionally, the AFLP method can be a costly technique since it requires an expensive software package to analyze a large number of AFLP patterns.



Schematic representation of the Amplified Fragment Length Polymorphisms (AFLPs) method. Genomic DNA is digested by restriction enzymes (1) and adapters are ligated to the restriction fragments (2). By using primers with selective nucleotides at the 3´-end, only a subset of the ligated fragments is amplified (3). Species (A and B) are identified by running the amplified products on a conventional electrophoretic gel (4).

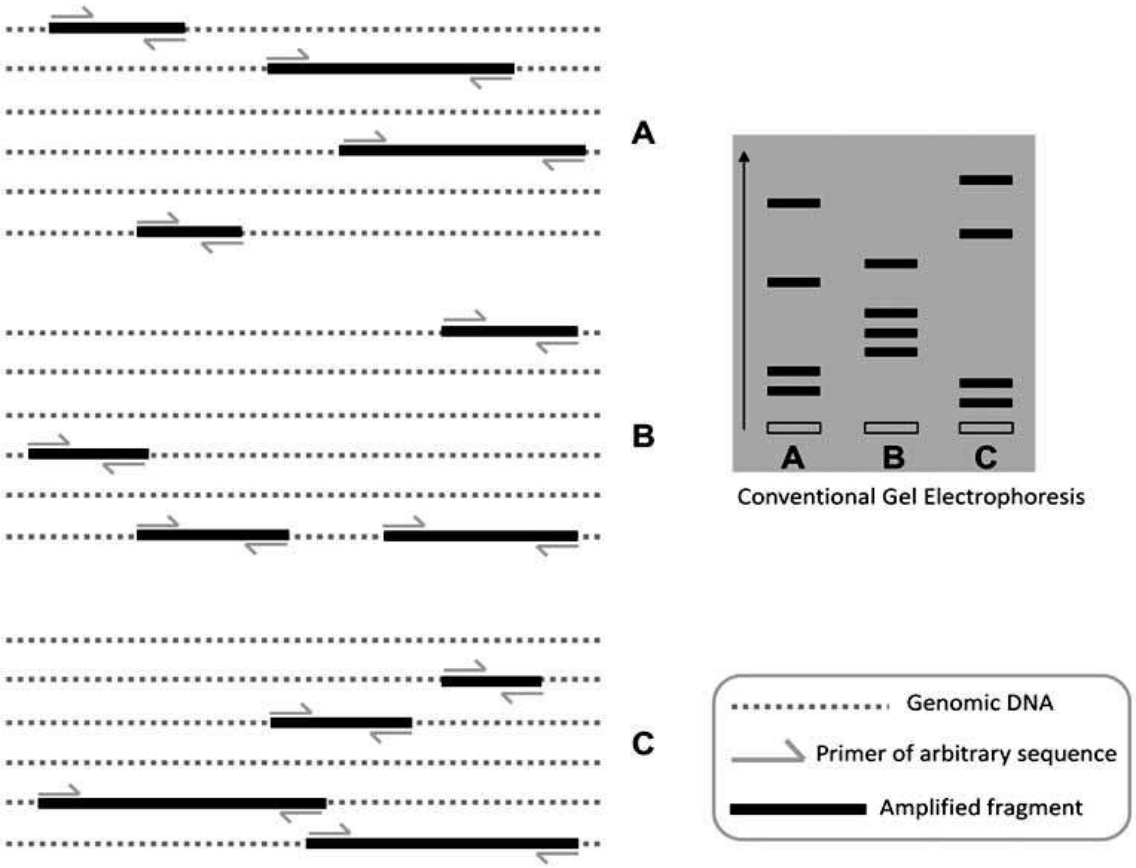
Random Amplified Polymorphic DNA (RAPD).

- **RAPD profiles are generated by the random PCR amplification of DNA segments using short primers of arbitrary nucleotide** sequence of usually 9 or 10 nucleotides long .
- These primers hybridize with sufficient affinity to different genomic regions at low annealing temperatures. Amplification products are generated when two RAPD primers anneal within a few thousand bases of each other in the proper orientation.

- Each species is identified by a specific banding pattern in an electrophoretic gel or similar technique resulting from the different genomic location of primer-binding sites .
- This technology is also known as arbitrarily primed-polymerase chain reaction(APPCR) and has been successfully used in a number of studies .

- **The RAPD method does not** require prior sequence information for PCR primer design but is extremely dependent on variations in laboratorial conditions (such as template DNA concentration, PCR and electrophoretic settings, etc), needing carefully developed laboratory proto- cols to be reproducible. An imperfect hybridization between the primer and the target site may result in a completely different banding profile.

- **The RAPD method, as well as other fingerprinting techniques, generates** results that can be difficult to interpret in cases where biological materials from different species are present in the sample (for instance, in some food products or in biological material from an individual infected with parasites). Another disadvantage is the need of purified DNA of high molecular weight.



Schematic representation of the Random Amplified Polymorphic DNA (RAPD) method. Species (A, B and C) are differentiated by the annealing of a single primer of arbitrary nucleotide sequence to different genomic regions. The amplified segments of DNA are separated and visualized in a conventional electrophoretic gel.

Conventional PCR.

- In recent years, a number of approaches based **on conventional PCR techniques have been described as a tool for species identification**
- Usually, **a conventional PCR-based method consists in the design of PCR primers that will only originate an amplification product in the presence of DNA from the target species.**
- The process of designing species-specific primers is now straight forward due to the vast number of genomic sequences available and software programs that assists in primer designing.

- A drawback of this technique is that it does not provide information about the presence of biological material from species that are not the target of the primers. A positive result may give an idea about the presence of a particular species, but a negative result gives no information about the origin of the sample (except that it does not belong to the species for which the assay has been designed for).

- To avoid this problem, prior sequence knowledge is necessary to derive specific primers for all species suspected to be present in the sample. An additional disadvantage is the need of performing an electrophoresis after the PCR to verify the amplification success of expected target sequences.

Real-time PCR.

- The basic goal of real-time PCR is the detection of a specific DNA sequence in a sample by measuring the accumulation of amplified products during the PCR using fluorescent technology. An important benefit of this method is the capability to quantify the starting amount of a specific DNA sequence in the sample (this approach is also known as quantitative PCR).

- The ability to monitor the progress of DNA amplification in real time depends on the chemistries and instrumentation used.
- Generally, chemistries consist of special fluorescent probes that must associate a fluorescent signal to the amplification of DNA.
- Several types of probes exist, including DNA-binding dyes like ethidium bromide, hydrolysis probes (5′-nuclease probes), hybridization probes, molecular beacons, PNA light-up probes, etc.

- The real-time PCR has the advantage over conventional PCR-based identification systems of working without post- PCR handling, with a minimised risk of carryover contamination in the laboratory.
- It also offers an increased sensitivity by permitting the discrimination of spurious PCR amplifications from non-target DNAs and is a relatively fast genotyping method, with some platforms affording high- throughput automation.

- Most severe disadvantages of **real- time PCR methods** **are the incompatibility of certain plat- forms with some fluorescent dyes, the restricted multiplex capability and the high cost of most reagents and instrumentation.**

Sequencing of PCR products.

- The DNA sequencing analysis **is currently the most used method for molecular species identification**. The advent of rapid and cost-effective PCR-linked DNA sequence analysis has circumvented the need for screening of genomic libraries and cloning of DNA fragments.

- The identification **is achieved by comparing the sequence of a genomic region found in the target sample with a comprehensive reference database**. Ideally, the structure of the DNA region to be analysed **must consist of a variable sequence** (informative enough to discriminate species) flanked by highly conserved regions (ideal to design universal PCR primers that amplify in a large number of species).

- A common way to assign a particular sequence to its species of origin is to perform a **BLAST search on the vast GenBank sequence database**.
- However, care must be taken when assigning the questioned sequence to the species with the highest similarity, because several gaps and false sequences are known to be present in these databases.
- Moreover, this approach does not provide any information and can lead to false identifications if the target sample belongs to a previously uncharacterized species.

- An advantage of sequencing ribosomal RNA (rRNA) genes **is the presence of conserved region** (for instance, 18S rRNA in eukaryotes and the 16S rRNA in prokaryotes) adjacent with highly variable segments (such as the internal transcribed spacers) allowing the resolution of relationships among both distantly and closely phylogenetic related spe- cies, respectively.

- The list of studies using **mtDNA cytochrome b gene for species identification is extensive .**

- This gene shows a high level of congruence with species limits and can be amplified in several vertebrate species under standard conditions by using a single pair of universal primers .

- Recently, a DNA-based barcoding system for all animal species has been proposed based on 650 to 750 bp of the **mtDNA cytochrome c oxidase (COI) gene** .

- **The "DNA barcoding"** concept is not an entirely new idea but, for the first time, it has been proposed to work at large-scale under well-defined standardized protocols .
- It has been projected both to assign unknown individuals to species and to facilitate the species-discovery process .
- The approach is controversial, with critics questioning both the method and its applications .
- Most important concerns are related to the use of a single gene in delineating and identifying species and the extent of separation between intra- and interspecies variations .
- Moreover, the COI system is obviously limited to eukaryotic species with mtDNA.

- A general drawback of DNA-sequencing approaches is that, in order to provide enough information for a secure discrimination, most of them rely on the sequencing of large DNA regions, usually over 300 bp .
- The PCR amplification of such large regions is difficult to obtain from samples with low quality and/or low amounts of DNA. The total amount of DNA available for analysis can be increased by performing a whole genome amplification prior to the sequencing.

DNA microarrays or DNA chips.

- It consists of small glass microscope slides, silicon chips or nylon membranes containing a large number of immobilized DNA fragments arranged in a regular pattern. A DNA microarray provides a medium for matching a reporter probe of known sequence against the DNA extracted from the target sample of unknown origin.

- Probes can include synthetic oligonucleotides, amplicons or larger DNA/RNA fragments selectively spotted or addressed to individual test sites in the microarray. The microarray is scanned or imaged to obtain a complete hybridization pattern generated by the release of a fluorescent, chemiluminescent, colorimetric or radioactive signal asso- ciated with the binding of the probe to the target DNA sequence .

- A DNA microarray built with species-specific DNA sequences can be used for identifications purposes .
- For instance, the DNA extracted from the target sample could be labelled with a specific fluorescent molecule and hybridized to the microarray DNA.
- A positive hybridization is detected with appropriate fluorescence scanning/imaging equipment (fluorescent spots are visualized). The DNA microarray hybridization methodology can also be directed for the screening of samples for species-specific single nucleotide polymorphisms (SNPs).

- Advances in printing technology have enabled the production of microarrays containing hundreds of thousands of probes (high-density microarrays may have up to 106 test sites in a 1-2-cm2 area), revealing the potential to achieve sensitive and high-throughput species identifications .

- PNA probes and molecular beacons can also be applied to the microarray technology for a rapid and large-volume systematic analysis of genetic information. **Nevertheless,**

- DNA microarrays require specialized robotics and imaging equipment that generally are not available in most laboratories.
- Advanced bioinformatics tools are also necessary to reduce the complex data into useful information.

Questions