Species Identification Using DNA Loci

4

ADRIAN LINACRE AND SHANAN S. TOBE

Contents

4.1 Introduction 61
4.2 Definition of a Species 62
4.2.1 Nomenclature 63
4.3 Previous Work in Species Identification 64
4.4 Genetic Polymorphisms 65
4.5 Mitochondrial DNA 67
4.6 Cytochrome *b* Gene 72
4.7 Phylogeny Trees 74
4.8 Species Identification Using the Cytochrome *b* Gene 80
4.8.1 Case Example 82
4.9 Other Mitochondrial Gene Loci 84
4.9.1 Bar Code for Life 84
4.10 Single Nucleotide Polymorphisms 85
4.11 Mixtures 86
4.11.1 Case Example 88
4.12 Problems with Mitochondrial DNA Markers 90
4.13 The Future of Species Testing 90
References 90

4.1 Introduction

The aim of this chapter is to describe the DNA tests that are used currently to identify an unknown sample as being of a particular species. This is a prime question in wildlife investigations, where the question "what species did this sample come from?" is frequently asked.

Species identification is a requirement for a large number of different reasons. Customs officers require the ability to identify species (or products derived from those species) being poached or traded contrary to national and international legislation. Animal interference after death can be confused with pre-mortem injuries, which can arouse suspicion of violence prior to death [1]. Rodents [2] and canids [3] frequently alter remains. Other needs include determining the species of origin of bloodstains [4]; identification of

the components of meats (for religious purposes or to identify poached species that are out of season or endangered, or where a quota has been exceeded) [4]; and identification of human and animal components of commingled remains from mass disasters, fires, cremations, and domestic crimes [5].

It is important to be able to clearly identify animal species to determine population numbers, diversity, and distribution. Non-invasive sampling methods for species identification, such as identification from scat [6], are required. A sensitive, quick, reliable, and accurate means is needed to clearly identify animal species present in samples [4], including mixtures. Although new species are continually being discovered, the number of species disappearing is at a far greater rate. For this reason, conservationists are interested in preserving and protecting native and endangered species and stopping the introduction of invasive species. International legislation (stipulated by CITES) covering the illegal trade of endangered species stipulates which species are protected. Chapter 2 outlined the role of CITES in the enforcement of trade of protected species with a range of species listed in CITES Appendices I, II, and III. National legislation within many countries may stipulate the protection of listed species; for instance, this includes particular bird species in the United States and deer species in India. A prime question asked of a forensic science laboratory is therefore to determine which species is present and if this species is protected by either international or national legislation.

To be able to enforce such legislation, it is necessary to be able to define a species.

4.2 Definition of a Species

There are a number of definitions of a species, and currently the two main definitions are biological species and phylogenetic species [7, 8]. The biological species definition is based on gross morphological features, whereas the phylogenetic species definition represents the relationships between organisms as revealed by their evolutionary history [9].

An example of the definition of a species from the *Oxford English Dictionary* is "a group of living organisms consisting of similar individuals capable of exchanging genes or interbreeding." This corresponds with the biological definition of species. To expand, the biological definition of species assumes that each species is reproductively isolated (genes are passed within a species but not beyond) [7, 10]. One species cannot breed with a different species and produce viable young. The problem with any definition is that there are exceptions. This concept holds true when dealing with animal hybrids such as mules (which are infertile because horses and donkeys have different numbers of chromosomes, so the mule offspring will have an odd number) but runs into problems when dealing with hybrids of plants (which

are able to reproduce) [7]. However, wolves, coyotes, and dogs can mate with each other and share the same number of chromosomes [11]. Even though mating is possible between these species, it is precluded by social factors and size differences, meaning that mating is a rare occurrence [11].

The phylogenetic definition of species relies on genetic markers (RNA, DNA, or protein) thought to be characteristic of a species [7, 9]. This is being widely used in taxonomy, biodiversity, and evolutionary studies [8]. Even with the advent of DNA and sequencing technologies, there is still no consensus as to how many genetic variations constitute a separate species.

4.2.1 Nomenclature

Normally all the members of the same species share a similar physical appearance, although this may be different between the sexes. It was this physical appearance that led Carl Linnaeus to develop a binomial Latin name for all known species. His system assigns a one-word name to a genus, to which more than one closely resembling species may belong, and a two-word name, with the first word being the same as for the genus, to a species. This system of species identification based on physical appearance was satisfactory for many years, but the advent of genetics has altered the classifications of some species. Further, there are sub-species (varieties within a species) where there may be phenotypic differences but individuals are capable of interbreeding and producing viable young. A well-known example of how genetics has affected the taxonomy of a species is the giant panda (*Ailuropoda melanoleuca*), which was originally considered as a bear but subsequently grouped with the red panda (*Ailurus fulgens*) although it shares much of its phenotype, and also its behavior and method of reproduction, with members of the bear family. The giant panda is currently placed within the bear family (*Ursidae*).

Tigers are used here as an example to demonstrate species and taxonomy. There are five currently living (extant) sub-species of tiger. The species name for tiger is *Panthera tigris* (note a species name is written in italics). The five sub-species are *P. tigris tigris* (Bengali), *P. tigris altica* (Siberian), *P. tigris corbetti* (Indo Chinese), *P. tigris amoyensis* (Amoy), and *P. tigris sumatrae* (Sumatran). All tigers are members of the Panthera genus. This genus includes lions (*Panthera leo* and *Panthera persica*), leopards (*Panthera panthera*), pumas (*Puma concolor*), and other large cats. All members of the Panthera genus belong to the Felidae family. This also includes the domestic cat (*Felis catus*). Going one category higher, all the Felidae belong to the order Carnivora, which also contains dogs (family Canidae) foxes (family Vulpes), and bears (family Ursus), among others. All the carnivores are members of the class Mammalia.

An example is given in table 4.1 of the classification of wolf, dog, fox, cat, and human. All are members of the class Mammalia and all are members of

Table 4.1 A Demonstration of the Increased Sequence Variation in the Cyt b Gene as Species Become More Distant Taxonomically

	Wolf	Dog	Fox	Cat	Human
Phylum	Chordata				
Class	Mammalia				
Order	Carnivora				Primate
Family	Canidae			Felidae	Hominid
Genus	*Canus*	*Canus*	*Vulpes*	*Felis*	*Homo*
Species	*lupus*	*lupus*	*vulpes*	*catus*	*sapiens*
Sub-species		*familiaris*			

Note: The first column lists the terms used in taxonomy starting at the bottom with the species name and ending with the phylum Chordata, into which all animals with a central nervous canal at the back (the spine in mammals) are grouped. There are two classifications higher, and those are kingdom and domain.

the order Carnivora, except humans, which are in the order Primate. Wolf, dog, and fox are members of the family Canidae, followed by genus Canus. Dog is here classified as a sub-species of wolf (*Canus lupus*), with dog as *Canus lupus familiaris*.

As the phenotype of the animal is affected by the genes it carries, it is not surprising that there is a relationship between the genetics and physical appearances of most species. The creation and maintenance of a species (speciation) are beyond the scope of this chapter; however, in general terms, when populations become isolated for a long period of time, chance mutations in the DNA result in genetic variations. These alterations to the DNA will only be shared and perpetuated by the members of the isolated group, and hence over time a new form, or species, may result.

4.3 Previous Work in Species Identification

If gross morphological characteristics are present (full or large intact portions of remains), it may simply be a case to identify the remains by microscopy or oesteology (analysis of the skeleton) for species identification so that no further examination is needed. However, if an animal is killed for food or sport, identifying characteristics may be intentionally removed [4], making morphological methods unsuitable [12], and therefore other methods of evaluation are needed [5]. Skins or pelts can be identified from microscopic analysis of the hairs, as the hairs of many animal species have distinct morphological characteristics. However, forensic science laboratories in Asia often get powdered samples with no morphological characteristics for identification [13], rendering this type of analysis unusable.

Other forms of examination, such as scanning electron microscopy (SEM) or gas chromatography coupled with mass spectroscopy (GC-MS), can be used. An inductively coupled plasma mass spectrometer (ICP-MS) can also be used, especially if the area of origin is in question, as captive bred animals may have different concentrations of elements in their tissues as a result of a differing diet from wild types (see chapter 6). Such details may be absent if the remains are as powder or ornaments. In such cases it may be necessary to determine the species by molecular means.

Traditionally, antigen–antibody reactions were used to identify species [14–16] and are indeed still used in many laboratories today [17]. There are several problems with this method of testing. One problem with this technique is that an antibody must first be produced and isolated [14, 15, 18]. This requires a large amount of starting material so that sufficient proteins are extracted for identification [15]. Many proteins lose their biological activity soon after death and can be subject to modification in different cell types [4]. Some antiserum, such as horse (H type), work only in a narrow range of concentrations and, under unstable conditions, can give rise to multiple precipitates [14, 18]. Other commercial antisera often suffer from low titre and cross-reactivity [14, 18]. It may be necessary to titre the batch of antibody prior to use [14].

Since creating an antibody requires a lot of time and effort, and a different antibody is needed for each animal species being tested, new molecular techniques are being developed. In addition, samples that may require analysis include traditional East Asian medicine, bone samples, horns and hairs, sculptures, and powdered remains, which will have insufficient morphology for analysis and may not contain any, or enough, proteins for analysis. However, DNA may be found in all these types of materials. Polymorphisms within this DNA allow for DNA-based techniques to be used to identify species.

4.4 Genetic Polymorphisms

There is a small amount of DNA variation even within a species, known as intraspecies variation (see box 4.1 for comments on DNA). It is estimated that the size of the human genome is about 3.2 billion base pairs. Any two humans taken at random are estimated to share at least 99.8% of the 3.2 billion bases. Humans are thought to have separated from their nearest genetic relative, the chimpanzee, about 6 million years ago. The amount of genetic similarity is greater than 98% and even greater at regions of DNA that encode proteins or are gene related. The differences between humans and chimpanzees are termed interspecies variations.

Box 4.1 DNA—The Basics

DNA is a very simplistic molecule essentially consisting of two extremely long chains, each comprising only a series of four bases. These bases are termed A, G, C, and T. The two chains are wrapped around each other with particular pairing so that a G and C are always across from each other, as are A and T—this forms the famous double helix. Each pair is termed a base pair, and the human genome is estimated to be about 3.2 billion base pairs in total. The human genome is not particularly large and other mammalian species have genomes of comparable size. Plants such as some tree species have genomes that are orders of magnitude larger than the human genome. The 3.2 billion base pairs of the human genome are grouped into 23 sections, termed chromosomes. It should be noted that a chimpanzee has 24 chromosomes and the second largest human chromosome is a fusion of two smaller chimpanzee chromosomes. All the chromosomes are located in the central body of a cell termed the nucleus. In the outer area of a cell (the cytoplasm) there exist small bodies (mitochondria) whose function is to produce energy. Mitochondria have their own DNA, being a circular helical chain.

The function of the DNA is to produce proteins and related biological material. The human genome contains about 25,000 genes. Only about 1.5% of the entire genome is related to this type of biological activity. Much of mammalian DNA appears to have little known function and separates the regions that code for a protein. Over time, changes in the DNA sequence can occur. These changes occur slowly, such that closely related organisms will share much of their DNA if they have a common ancestor. Two humans taken at random share about 99.8% of their DNA, with only 1 base pair being different out of every 1,000. When comparing chimpanzee DNA to that of humans, there is approximately 98.8% homology. As human and species that are related genetically more distantly are examined, the degree of homology reduces. The DNA sequence in a gene may be conserved and varies much less than DNA that is non-gene related. The comparison of the DNA sequence of a gene for two members of the same species will be almost exactly the same. If there are any differences, then this is termed intraspecies diversity. The same DNA sequence, when compared between two different species, may show greater variation, termed interspecies variation. Greater interspecies variation may occur when two species are further apart genetically.

It would be expected that two species that have diverged more recently from a common genetic ancestor will share greater amounts of their DNA compared to a species with a more distantly related common ancestor. Dogs are thought to have been domesticated from wolves about 12,000 years ago

[19]. This is a very short space of time in genetic terms, and hence there is a high degree of homology. Wolves and other members of the Canus family show greater DNA sequence diversity.

In order to make use of the sequence diversity for species testing, it is necessary to first isolate and characterize the sections of DNA that are variable. This may almost certainly require some knowledge of the DNA of the species in question. In some cases this knowledge may not be available.

DNA-based tests that require no prior sequence information can be used, such as random amplification of polymorphic DNA (RAPD), amplified fragment length polymorphism (AFLP), and restriction fragment length polymorphism (RFLP). A comparison of these methods is shown in table 4.2. RAPD, AFLP, and RFLP have been applied to identify animal components [20–31]. The problem and limitation with these whole genome techniques are that there are always issues with reproducibility, and when complex mixtures of two or more species are to be detected, interpretation of results may be difficult due to overlapping restriction patterns that may be generated [20]. It is also difficult to produce accurate reference databases for comparison purposes. It is therefore not only impractical but also logically nonsensical to examine the whole genome when developing a DNA-based species test. It is preferable, then, to use the regions of DNA that show no, or little, intraspecies variation but sufficient interspecies variation for identification purposes. To use known genetic loci, there must be prior knowledge of the genome of the animal being studied. This information is often provided from studies compiling phylogenetic trees attempting to link the ancestry of different animal species or to trace the movement and development of new species. Many of these genes are located on the mitochondrial genome.

4.5 Mitochondrial DNA

The DNA regions used in taxonomy and species testing need to show sufficient interspecies variation to be able to distinguish closely related species. Further, the samples are often old, highly degraded, and powdered. For this reason, and reasons described in the following, the DNA regions used for species testing are on the mitochondrial genome.

The mitochondrial genome is a circular loop of DNA housed within the mitochondrion. Mitochondria are present in the cytoplasm of almost all cells. The number of mitochondria varies between cell types, as the role of the mitochondria is to generate chemical energy, adenosine triphosphate (ATP), by the process of oxidative respiration, in which oxygen is converted to carbon dioxide [46]. Muscle cells contain more mitochondria, and hence more mitochondrial DNA, compared to most skin cells.

Table 4.2 A List of Some of the Current Techniques Used for Species Identification. A Brief Description Is Provided as Well as Advantages and Disadvantages of Each Technique.

Technique	
Advantages	Disadvantages
RAPD: Uses short, random oligonucleotide primers to amplify arbitrary segments of DNA to produce a band pattern specific to an individual or species.	
• Gives "fingerprint" of entire genome	• Difficult to reproduce results within and between laboratories
• Cheap	• Cannot produce a database
• Fast	• Cannot reliably separate mixtures [24]
• No prior gene sequence knowledge is needed [32]	• Even among the same species variations exist between different samples [33, 34]
	• Quality and quantity of template will affect the band pattern [33, 35]
	• Buffer, dNTPs, and primer concentration all influence band patterns [35, 36]
	• Major bands shift with degraded DNA [33]
	• Efficiency of primers will affect results [36]
RFLP/AFLP: The use of restriction enzymes to break DNA into fragments. Identification is based on the size and pattern of the resulting band patterns.	
• Generates species-specific band patterns	• Requires large amounts of DNA if not coupled with PCR
• Can be coupled with universal or species-specific primers and PCR	• Several restriction enzymes are needed
• No prior gene sequence knowledge is needed (although it is beneficial in order to choose which restriction enzymes to use)	• Cannot reliably separate mixtures of restriction patterns [20, 37]
	• Expensive [12]
	• Laborious and time consuming [12]
	• Variation of restriction sites within species [37, 38]
	• Cannot separate closely related species [37, 39]
Sequencing: Amplification of a segment of DNA for which the base sequence is then determined. Has many different names for the same process: barcoding [40–43], short mitochondrial informative regions (SMIRs) [44], and forensically informative nucleotide sequencing (FINS) [4]. Single nucleotide polymorphisms (SNPs) would also be included in this category as they are an analysis of sequence variation of a smaller scale.	
• Can analyze a wide range of species if using universally applicable primers	• Cannot separate mixtures
	• Degraded samples may not generate enough sequence data for identification
	• Susceptible to contamination if using universal primers

Table 4.2 A List of Some of the Current Techniques Used for Species Identification. A Brief Description Is Provided as Well as Advantages and Disadvantages of Each Technique (Continued).

Technique	
Advantages	Disadvantages
Species-specific primers: Isolate SNPs specific to a species and primers are designed based on those SNPs. Products will only be obtained for the species for which the primers were designed.	
• Able to separate complex mixtures [45]	• Prior sequence knowledge is needed to design primers
• It is possible to add additional species as needed	• Only those species for which primers have been designed can be identified
• No need for post-PCR digestion or sequencing	
• Can use multiple genes simultaneously	
• If species-specific primers are coupled with labeled universal primers, can be quite cheap	

Reasons for using mitochondrial DNA rather than DNA within the nucleus include:

- Multiple copies: Each mitochondrion contains its own DNA, with many copies of the circular mitochondrial DNA in every cell. It is thought that each mitochondrion contains between 1 and 15, with an average of 4 to 5, copies of the DNA [47] and there are hundreds, sometimes thousands, of mitochondria per cell. The result is that there are many thousands of copies of the mitochondrial DNA in every cell. This compares with only two copies of nuclear DNA.
- Better protection: The mitochondrion also has a strong protein coat that protects the mitochondrial DNA from degradation by bacterial enzymes. This compares to the nuclear envelope that is relatively weak and liable to degradation.
- Higher rate of evolution: DNA alterations (mutations) occur in a number of ways. One of the most common ways by which mutations occur is during DNA replication. An incorrect DNA base may be added; for example, a C is added instead of a G. This creates a single base change, or polymorphism, resulting in a new form. These single base mutations are rare but occur once every 1,200 bases in the human genome. Single nucleotide polymorphisms are created more commonly than are observed as the errors that occur during DNA replication are corrected (proofread) by an enzyme that exists within the nucleus. This error correcting enzyme does not exist in

the mitochondria [48], and hence errors made when the circular loop of DNA replicates are not corrected. The result is that the rate of change, or evolutionary rate, of mitochondrial DNA is about five times greater than nuclear DNA [48]. This is important in species testing, as even species thought to be closely related may in time accumulate differences in the mitochondrial DNA but show little difference in the nuclear DNA.

- Maternal inheritance: A further reason for the use of mitochondrial DNA in species testing, and in forensic science, is its mode of inheritance. Mitochondria exist within the cytoplasm of cells, including the egg cells. Spermatozoa do not normally pass on mitochondria and only pass on their nuclear DNA. The resulting embryo inherits all its mitochondria from its mother [49–51]. The result is that mothers pass on their mitochondrial DNA type to all their offspring, but only the daughters will pass on the mitochondrial DNA to the next generation. Mitochondrial DNA is therefore passed from generation to generation down the maternal line. There are exceptions to this rule, as demonstrated by mussels of the genus *Mytilus*, which have biparental inheritance of mitochondria [52]. Also, unlike nuclear DNA, where there is a shuffling of the chromosomes at every generation, the mitochondrial DNA does not recombine with any other DNA type and remains intact from generation to generation [50, 53].

The role of DNA is to encode protein and RNA molecules, and the mitochondrial DNA is no different. It possesses 37 genes, encoding 22 tRNA molecules, 2 rRNA molecules, and 13 proteins involved in respiration. The number of bases that comprise the human mitochondrial DNA is about 16,570 base pairs [54]. All mammalian mitochondrial DNA is very similar, with the order and position of the genes being the same. The general structure of the mitochondrial DNA is shown in figure 4.1. There is a site on the mitochondrial DNA where the two strands unzip as part of the replication process, and by convention a base in the middle of this area is called base 1. The DNA bases are numbered sequentially, so in humans the base to one side of base number 1 is base 2, and to the other side it is base 16,570 (note some human mitochondrial DNA types are only 16,569 in length).

The region around base number 1, where the DNA unzips, is the only part of the mammalian mitochondrial DNA that does not encode a protein. As this area is non-coding, it is the one area on the mitochondrial DNA that can develop DNA mutations without affecting the role of the DNA. Greater intraspecies variation occurs at this region compared to the coding regions of the mitochondrial DNA. The DNA bases on either side of base number 1, extending to base 576 in one direction and starting from base 16,024, are called

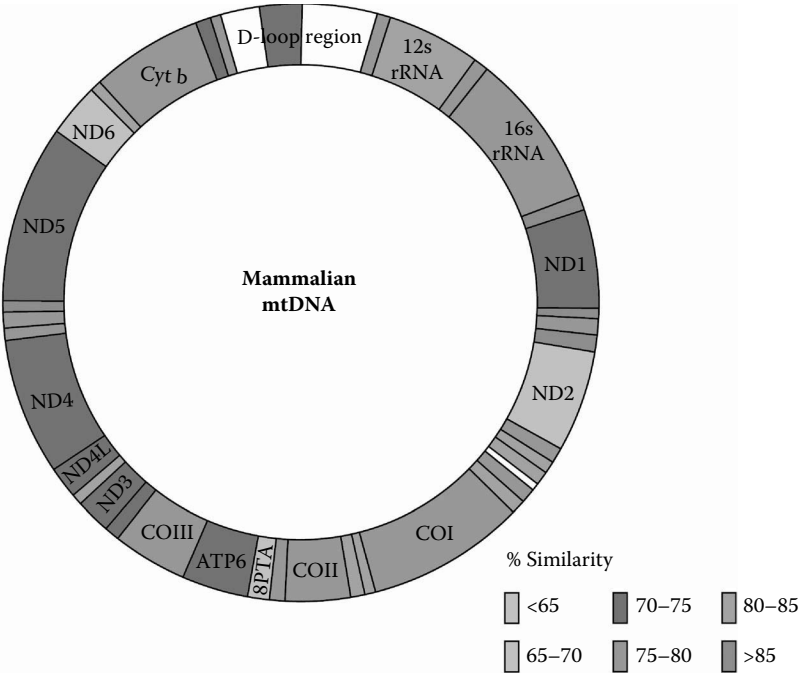

Figure 4.1 **(See color insert following page 114.)** A schematic view of the mammalian mitochondrial genome. It is a circular loop of double-stranded DNA. Encoded along the loop are 37 genes; these are of various sizes in length and produce proteins or RNA molecules required by the mitochondrion. Base number 1 is at 12:00 as viewed and the bases are numbered sequentially. The bases on either side of base 1 are the only non-coding sections of the DNA, termed hypervariable regions 1 and 2. Next to hypervariable region 1 is the cytochrome *b* gene running from base 14,756 to 15,896 on the reference human mitochondrial genome [54]. The varying colors denote the degree of similarity of the DNA sequences between mammalian species.

the hypervariable regions. These regions, hypervariable region 1 from 16,024 to 16,365 and hypervariable region 2 from 73 to 340, are used in human identification and forensic science to identify a particular individual human. The amount of intraspecies variation precludes their use in species identification.

Gene sequences encode a particular protein, and therefore there is selection pressure against change in the DNA sequence. This conservation of sequence is very important as it preserves the structure and function of the protein produced by the gene. Fewer DNA alterations occur within these encoding regions, resulting in little intraspecies variation. There is also little interspecies variation, as the protein fulfils the same task in a blue whale as it does in a harvest mouse. Proteins are made of amino acids, and amino acids fall into four major categories (acidic, basic, polar, and nonpolar). A mutation can occur that alters, for instance, a basic amino acid for another basic

amino acid without a loss of function to the protein. Additionally, a mutation can occur within a gene sequence without any alteration to the amino acid; for instance, the amino acid proline is determined by the three bases CCN, where N is any of the four bases (CCA, CCC, CCG, and CCT, all encode proline). This is an example of how the genetic code is redundant and the third base of the three that encodes an amino acid is often silent. Variation in the DNA sequence of a gene at this position can occur without adversely affecting the protein.

One gene on the mitochondrial DNA in particular has become a tool in species identification, and that gene is called the cytochrome *b* gene.

4.6 Cytochrome *b* Gene

The cytochrome *b* gene is 1,140 bases long and in humans is positioned between 14,756 to 15,896 on the reference human mitochondrial genome [54]. The DNA base number does vary between different mammalian species, but in all cases the cytochrome *b* gene is situated close to one side of hypervariable region 1. It encodes a protein that is 380 amino acids in length [55]; this length is invariable within mammalian species. A schematic diagram showing the position of the cytochrome *b* gene is shown in figure 4.2. Due to the function of the protein, some regions of the amino acid sequence show little variation and other regions have the possibility of greater variation.

Cytochrome *b* is one of 10 proteins that make up complex III of the mitochondrial oxidative phosphorylation system. It is the only one of these proteins encoded by the mitochondrial genome. The protein spans the membrane of the mitochondria and is thought to be involved in electron transfer. Cytochrome *b* crosses the membrane at 8 positions, and these transmembrane spanning amino acid regions show conservation between mammalian species. There is greater variation away from these active regions of the protein. At these less active regions, amino acid variation as described previously can occur. Small variations in the amino sequence encoded by the gene allied with small variation at the third redundant base of some DNA sequences encoding an amino acid lead to variation of the DNA between closely related species.

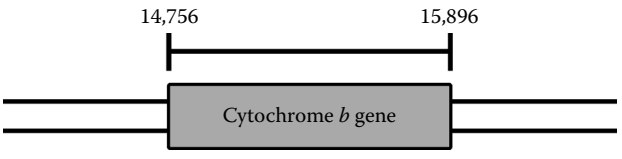

Figure 4.2 A schematic of the cytochrome *b* gene. The gene is close to HV1 on the mitochondrial DNA from 14,756 to 15,896 in humans.

Table 4.3 The Amount of Divergence Using the Entire Cytochrome *b* Gene for Five Mammalian Species Is Shown. This Is Shown as a Percentage Above the Diagonal and as the Number of Base Variations Below the Diagonal.

	Wolf	Dog	Fox	Cat	Human
Wolf		99.6	84.0	79.0	74.5
Dog	4		83.7	78.9	74.3
Fox	182	186		78.6	73.8
Cat	239	240	244		76.6
Human	291	293	299	267	

Adopted from [56].

The cytochrome *b* gene was originally used for taxonomic purposes to establish the phylogeny of species [56, 57]. These studies required the comparison of all or part of the DNA sequence encoding the cytochrome *b* gene. Two members of the same species would be expected to have the same DNA sequence, although small intraspecies variation may result in one or two base differences. Closely related species would be expected to have very similar DNA sequences as they diverged from a common ancestor, and hence the same ancestral species, in the recent past. As the time since a common ancestor increases, it is possible for greater variation to occur, leading to greater differences. If the differences between species are compared, it is possible to determine the amount of variation or similarity. The amount of variation both in DNA bases and as a percentage between the five previously discussed species at the cytochrome *b* gene is shown in table 4.3.

The degree of homology between the different species mirrors the taxonomic grouping shown in table 4.1. It would be expected that dog and wolf would show very little variation, as it is thought that they have a recent common ancestor (often abbreviated as RCA), more recent than the separation of the common ancestors of the wolf/dog with that of the fox. Figure 4.3 shows a graphical representation of the same five species.

The DNA sequence for the cytochrome *b* gene is known for a large number of mammalian species. DNA sequences are lodged with the DNA databanks as a repository of genetic information. Two of the largest databanks are the European Molecular Biology Laboratory (EMBL-Bank), maintained at the European Bioinformatics Institute in the UK (www.ebi.ac.uk), and GenBank, maintained by the National Center for Biotechnology Information in the United States (www.ncbi.nih.gov). In the EMBL-Bank alone there are more than 36 billion nucleotide entries from more than 150,000 organisms, and the number is growing continuously. The EMBL-Bank and GenBank joined together with the Japanese-based DNA Databank of Japan (DDBJ), and now information is shared among all three to form the International Nucleotide Sequence Database (INSD).

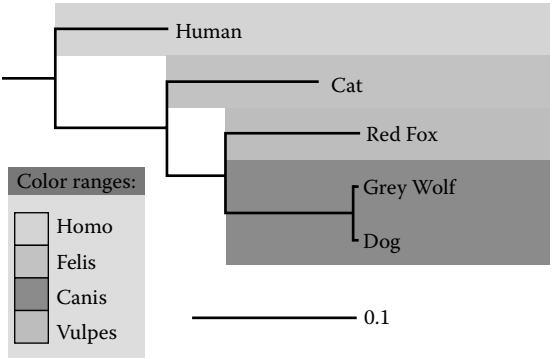

Figure 4.3 (See color insert following page 114.) A phylogenetic tree of the five mammals from table 4.1 analyzed based on the cytochrome *b* difference as described in table 4.2. The scale represents the number of differences between the organisms at the genetic level using the cytochrome *b* gene. The diagram was created using the Interactive Tree of Life [58].

Using the cytochrome *b* gene of the five species as an example (wolf, dog, cat, fox, and human), and when the DNA sequences are compared to the sequences of other closely related and unrelated species, a phylogenetic tree can be produced as in figure 4.4. This tree contains, in majority, mammalian species, but also included are species of bird, fish, and a reptile. The idea behind this diagram is to illustrate that species that are expected to be closely related genetically will group together. A DNA sequence from an unknown sample, when compared to the DNA sequences stored in the DNA databases, will either match perfectly with one of the known samples or will match to a high degree with one of the registered samples.

One problem with the DNA databases is that the DNA sequences registered have not been checked or verified. The accuracy of the sequence data is very much dependent on the supplier registering the data. With relatively recent advances in DNA sequencing there are fewer errors, although this was not the case with the original method of DNA sequencing. Additionally, many nonhuman sequences turn out to be human and are the result of contamination by a laboratory operator. To be confident of the sequence, the source of the original material should be verified.

4.7 Phylogeny Trees

Phylogeny trees are a graphical representation of the relationships between different organisms that share a common ancestor. The phylogeny tree represents in what order these species, or individuals, diverged from a common ancestor. There are many types of data that can be used to construct a phylogenetic tree.

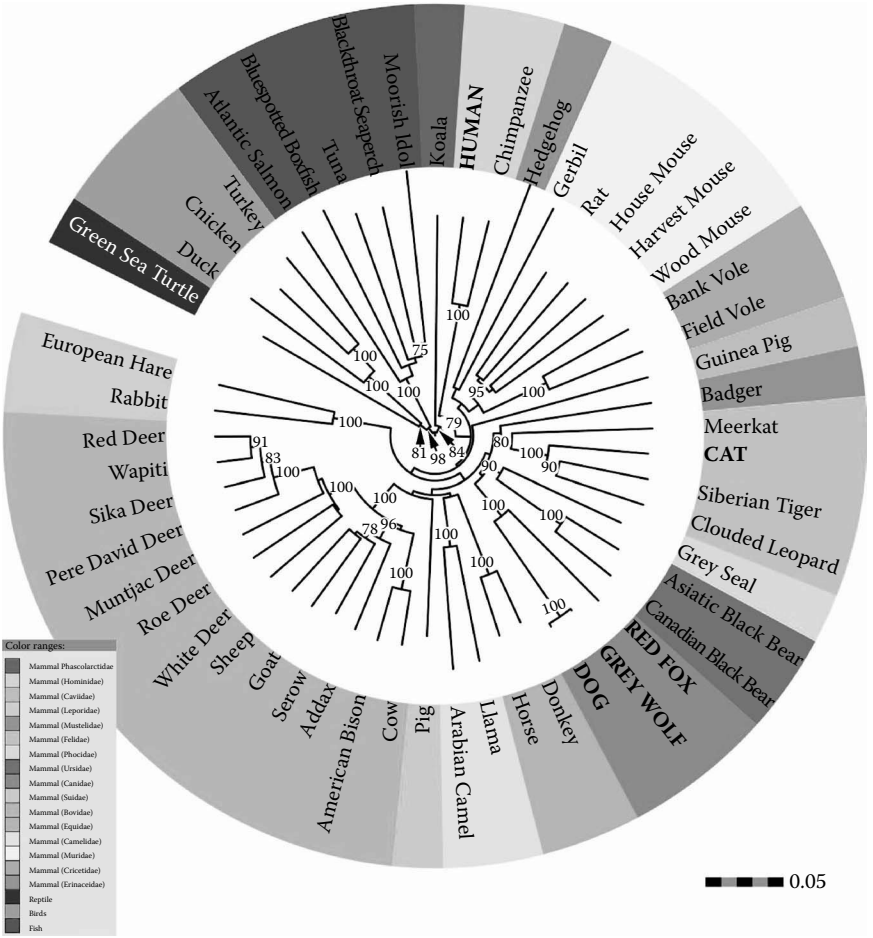

Figure 4.4 **(See color insert following page 114.)** A phylogeny tree displaying 52 species of animals, the original 5 animals in bold compared to 47 other animals, as produced based on alignment of the cytochrome *b* gene. Clades (groups of organisms with a common ancestor) are shaded to indicate relationships. Expected relationships are observed as dog and wolf, horse and donkey, cat species, and birds and fish are all observed grouped together. Bootstrap values greater than 75 are displayed. The scale represents the number of differences between the organisms in percent. Phylogeny trees and bootstrap values are explained in section 4.4 and box 4.2. Created using the Interactive Tree of Life [58].

For the purposes of this explanation, we will be using nucleotide sequence data from the cytochrome *b* gene on the mitochondrial DNA. For consistency, we will also limit our examples to the five mammalian species (cat, dog, fox, human, and wolf) that we have been using throughout the chapter.

Let us consider the alignment of cat, dog, fox, human, and wolf (figure 4.5). If we were to analyze each sequence against each of the others, we

```
Dog             ATGACCAACATTCGAAAAACCCACCCACTAGCCAAAATTGTTAATAACTCATTCATTGAC
Grey Wolf       ATGACCAACATTCGAAAAACCCACCCACTAGCCAAAATTGTTAATAACTCATTCATTGAC
Red Fox         ATGACCAACATTCGAAAGACTCACCCACTAGCTAAAATCGTAAACGACTCATTCATCGAC
Cat             ATGACCAACATTCGAAAATCACACCCCCTTATCAAAATTATTAATCACTCATTCATCGAT
Human           ATGACCCCAATACGCAAAATTAACCCCCTAATAAAATTAATTAACCACTCATTCATCGAC
                ******   ** ** **     **** **    *** *  * **  ********** **
```

Figure 4.5 A sequence alignment of the first 60 nucleotides of the cytochrome b gene on the mitochondrial genome for cat, dog, fox, human, and wolf. Homologous bases between all species are indicated with an * under them.

would find that, according to these sequences, the number of nucleotide differences between the species can be determined (table 4.4).

This information can then be displayed in a phylogenetic tree, of which there are different types, but the simplest form is the rooted tree. We will use the rooted tree for all of our examples. The root of the tree is the RCA for all of the species shown. From table 4.4 we can see that wolf and dog have less variation than do wolf and fox, which have less variation than wolf and human. From this we can conclude that wolf and dog have a more recent common ancestor than do wolf and fox. This can be converted to several different rooted phylogenetic trees (figure 4.6).

The terminal points (represented by the species names) are often referred to as terminal nodes. These represent the extant species data available, whereas the nodes where the data points join (internal nodes) represent a proposed common ancestor. Texts may refer to the extant species as operational taxonomic units. The aim of any phylogenetic tree is to represent the probable genetic distance between two extant species and represent the distance to the most recent common ancestor.

At first glance it appears that all three trees in figure 4.6 are different. In fact, figures 4.6A and B give the same information but are represented in a pictorially different way. The branches in A and B have been scaled according to the genetic distance between the species, and the scales represent the differences between the sequences in percentages. The only difference between

Table 4.4 The Amount of Divergence Using the Entire Cytochrome *b* Gene for Cat, Dog, Fox, Human, and Wolf Is Shown. The Number of Base Variations Is Shown.

	Wolf	Dog	Fox	Cat	Human
Wolf		4	182	239	291
Dog	4		186	240	293
Fox	182	186		244	299
Cat	239	240	244		267
Human	291	293	299	267	

Adopted from [55].

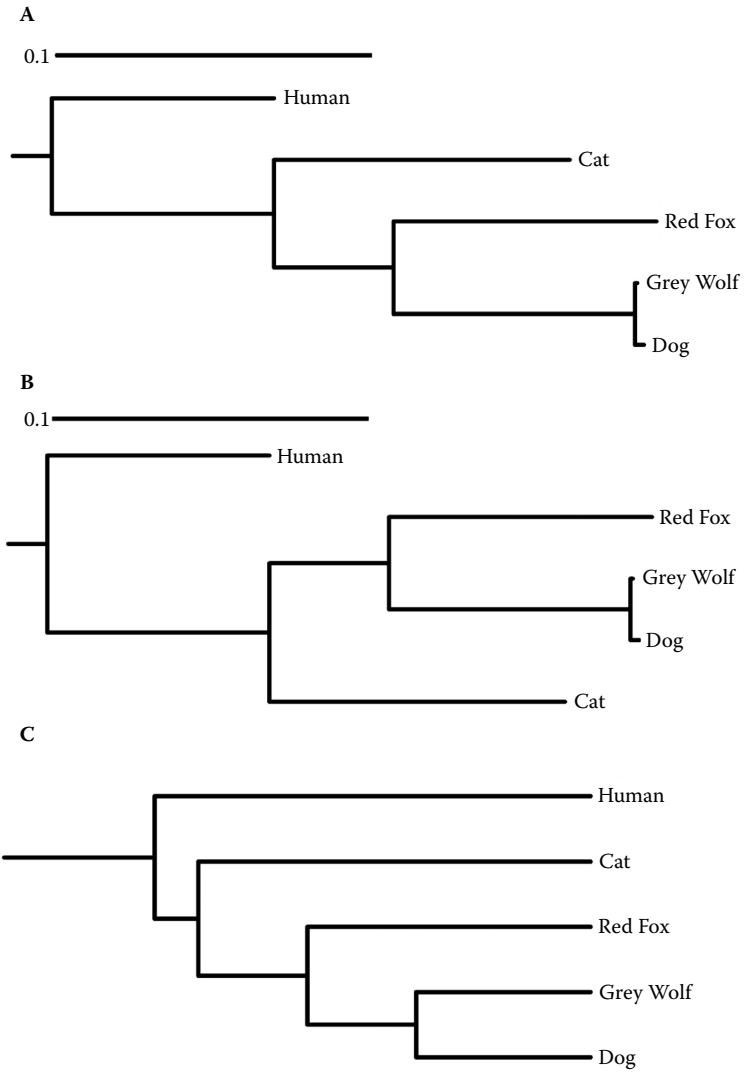

Figure 4.6 Three different rooted trees based on sequence data from the cytochrome *b* gene for cat, dog, grey wolf, human, and red fox. A and B are both rooted trees with the lengths of the branches indicating the percentage separation. C is a rooted tree with the branch lengths ignored, which gives less information than do trees A and B. Created using the Interactive Tree of Life [58].

the two is that the branch connecting cat and fox/wolf/dog has been rotated. Branches can be rotated around an axis without altering the table information. Therefore, red fox and dog/wolf can be rotated, but not human and red fox. Both trees make it obvious that the human sample diverged from the other species at an earlier date, followed by cat, then fox, and finally dog and wolf, which have the most recent common ancestor.

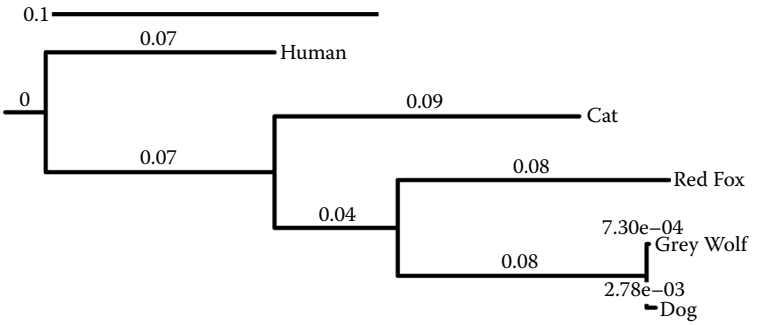

Figure 4.7 A rooted tree based on sequence data from the cytochrome *b* gene for cat, dog, grey wolf, human, and red fox. Branch lengths are displayed. Created using the Interactive Tree of Life [58].

Figure 4.6C, however, gives much less information. All the branches of the tree have been altered to allow the species to line up at the same point, and the figure is therefore nonscaled. This removes much of the information about how much change (how many bases) has occurred between the species, but it still reflects the order of divergence between the species. We can still see that dog and wolf diverged more recently than did cat and human.

The scale represents the percent difference between the species. An example of this is shown in figure 4.7.

It is important to realize that these diagrammatic relationships are based on current DNA data and the RCA is inferred. DNA data can be combined with, or used to support, archaeological remains. Despite the genetic data from sequences on the mitochondrial genome used in species identification and in particular to indicate the degree of diversity between two closely related species, there are few means of confirming whether the degree of genetic relatedness is supported by the data. One test that addresses this problem is called bootstrapping (box 4.2).

Box 4.2 Bootstrapping

Bootstrapping is a standard tool in phylogenetics. It is used to assess the reliability of a phylogenetic tree by taking the aligned DNA sequence data and reshuffling the columns. The reshuffling of the columns allows replacement, so the same column may appear multiple times and a column may not appear as the total number of columns will not change. An example of bootstrapping is shown in figure B4.1.

The idea of bootstrapping was first described by Efron [1] with the idea of re-sampling data many times to determine the degree of confidence in a hypothesis. Since then it has been a standard method in systematics and phylogenetics. It is very relevant to species testing, as a DNA sequence

Column number	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Dog	A	T	G	A	C	C	A	A	C	A	T	T	C	G	A
Grey Wolf	A	T	G	A	C	C	A	A	C	A	T	T	C	G	A
Red Fox	A	T	G	A	C	C	A	A	C	A	T	T	C	G	A
Cat	A	T	G	A	C	C	A	A	C	A	T	T	C	G	A
Human	A	T	G	A	C	C	C	C	A	A	T	A	C	G	C
	*	*	*	*	*	*				*	*		*	*	

Bootstrap Replicate 1

Column number	2	7	12	8	9	3	4	9	13	15	10	2	6	7	4
Dog	T	A	T	A	C	G	A	C	C	A	A	T	C	A	A
Grey Wolf	T	A	T	A	C	G	A	C	C	A	A	T	C	A	A
Red Fox	T	A	T	A	C	G	A	C	C	A	A	T	C	A	A
Cat	T	A	T	A	C	G	A	C	C	A	A	T	C	A	A
Human	T	C	A	C	A	G	A	A	C	C	A	T	C	C	A
	*					*	*				*	*	*		

Bootstrap Replicate 2

Column number	10	8	6	4	13	3	13	6	2	1	5	10	1	12	7
Dog	A	A	C	A	C	G	C	C	T	A	C	A	A	T	A
Grey Wolf	A	A	C	A	C	G	C	C	T	A	C	A	A	T	A
Red Fox	A	A	C	A	C	G	C	C	T	A	C	A	A	T	A
Cat	A	A	C	A	C	G	C	C	T	A	C	A	A	T	A
Human	A	C	C	A	C	G	C	C	T	A	C	A	A	A	C
	*		*	*	*	*	*	*	*	*	*	*	*	*	*

Bootstrap Replicate 3

Column number	15	3	4	7	11	12	3	1	2	7	9	8	9	8	5
Dog	A	G	A	A	T	T	G	A	T	A	C	A	C	A	C
Grey Wolf	A	G	A	A	T	T	G	A	T	A	C	A	C	A	C
Red Fox	A	G	A	A	T	T	G	A	T	A	C	A	C	A	C
Cat	A	G	A	A	T	T	G	A	T	A	C	A	C	A	C
Human	C	G	A	C	T	A	G	A	T	C	A	C	A	C	C
		*	*		*		*	*	*						*

Figure B4.1 How the randomized sequences are produced. The original data set of 15 columns of an aligned DNA sequence is shown. Homologous bases between all species are indicated with an * under them. The same data set is then shuffled based on reordering the columns. The number of columns remain the same (15), but the same column may appear more than once and some columns not at all. This reshuffling is performed between 500 and 1,000 times to obtain a degree of confidence in support of the phylogenetic tree. Real data sets would be the entire DNA sequence and therefore may be 500 bp or more.

from an unknown can be compared to known DNA sequences, aligned, and reshuffled. The higher the bootstrap value, the greater the confidence that the two sequences are homologous. This reshuffling with replacement can be as many as 1,000 times, and each of the reshuffles will generate a phylogenetic tree. If the phylogenetic tree is replicated in all 1,000 reshuffles, then a score of 100% is obtained; but if there is a difference in 20 out of the 1,000 reshuffles, then a score of 98% will be obtained. Ultimately a phylogenetic tree as shown in figure 4.4 is obtained.

Reference

1. Efron, B., Bootstrap methods: another look at the jackknife. *The Annals of Statistics*, 1979, 7(1): 1–26.

4.8 Species Identification Using the Cytochrome *b* Gene

DNA sequences from unknown samples are ultimately compared to the DNA databanks to determine if there is a match to any previously registered DNA sequence. The first step in this identification is the isolation of DNA from the unknown sample.

DNA can be isolated from trace materials with increasing success. Conventional processes work on the basis that the DNA molecule is highly negatively charged and therefore will bind to a substrate that is positively charged. Unbound cellular material can be removed by various solutions, and the DNA is finally isolated from the substrate once almost all the non-DNA cellular components have been removed. Commercial kits are now available that increase the sensitivity and reproducibility of the DNA extraction process.

The second stage is to take the trace amounts of DNA, target the cytochrome *b* region only, and amplify this section of DNA for analysis. The advent of the polymerase chain reaction (PCR) has transformed much of molecular biology, allowing trace amounts of DNA to be routinely analyzed. The PCR process is in essence a high-fidelity photocopier that will copy a specific region of DNA millions of times. The specificity of the copying process is reliant on DNA primers that bind either side of the DNA region to the amplified region. Currently, PCR products of about 800 bases or less in size are amplified routinely. The complete cytochrome *b* gene is slightly larger than the normal size of PCR products that are targeted, in which case smaller sections of the DNA can be amplified.

As the cytochrome *b* gene DNA sequence is very similar for all mammals, it is possible to produce primers that are universal for all the known mammalian species. The benefit of universal primers is that there is no need

for any prior information as to the species present in the unknown sample, as the primers will work on all mammalian DNA.

There are a few universal priming sites that have been developed and found to produce a PCR product from all known mammals. An example is the primer pair developed that amplifies a section at the start of the cytochrome *b* gene that is 402 bases in length [59] optimised by [13]. Although this represents less than a third of the entire cytochrome *b* gene, there is sufficient sequence diversity to distinguish almost all mammalian species.

Current technology permits rapid DNA sequencing of sections of DNA produced by the PCR process. The DNA sequences obtained can be compared to the DNA sequences registered at EMBL-Bank or GenBank. Sequence alignments by user friendly software programs such as Basic Local Alignment Search Tool (BLAST) can be performed. The output from the program ranks the DNA sequence matches in degrees of homology. If there is a perfect match between the unknown DNA sequence and the DNA sequence from a species in the databank with 100% homology, then this will be listed first. Those DNA sequences with lessening amounts of homology can be provided. An example of a section of a sequence alignment is shown in figure 4.8.

Figure 4.8 only shows an alignment over 50 bases rather than the entire cytochrome *b* gene, but even within this small section there is sufficient

```
Indian      CCAACGGAGCATCCATATTCTTCATCTGCCTATTTATTCATGTAGGACGA
Javan       CCAACGGAGCATCCATATTCTTCATCTGCCTATTTATTCATGTAGGACGA
White       CCAACGGAGCATCCATATTCTTTATCTGCCTATTCATCCACGTAGGACGC
Black       CTAACGGAGCATCCATATTTTTTATCTGCCTATTCATCCACATAGGACGC
Sumatran    CCAACGGAGCATCCATATTCTTCATCTGCCTATTTATCCACGTAGGACGA
unknown 1   CTAACGGAGCATCCATATTTTTTATCTGCCTATTCATCCACATAGGACGC
unknown 2   CTAACGGAGCATCCATATTTTTTATCTGCCTATTCATCCACATAGGACGC
Bovine      CCAACGGAGCATCAATGTTTTTCATCTGCTTATTTATGCACGTAGGACGA
            * ******* *** ** ** ** ****** **** ** **  *******
```

Figure 4.8 Sequence alignment of DNA sequences registered in EMBL from a small section of the cytochrome *b* gene. Represented are the five extant species of rhino, DNA from two unknown samples, and a cow as an out group. The unknown samples were from powdered remains sold as traditional East Asian medicines. Both the unknown samples have the same DNA sequence as the black rhino (*Diceros bicornis*). The asterisk denotes a DNA base position homologous to all samples. The bases that are highlighted denote those shared by the unknown with the black rhino only.

sequence diversity to be able to exclude the unknown samples as originating from any rhino species other than the black rhino.

The examples described relate to the analysis of traditional East Asian medicines and illegal trade in body parts. Nonhuman DNA testing can also be applied to more mainstream forensic science. Forensic wildlife DNA has been used in a number of countries for criminal investigative purposes, including Canada, Sweden, Austria, and Portugal. In the first instance, pet cat hair found on a murder victim was able to link a suspect to a murder scene and was a key point of evidence in the conviction of a murderer [60]. In Sweden, six cases involving dog and wolf hairs were reported [61]. In all cases it was shown that the suspect hair could not have come from dog hair associated with the suspects except in one case, where there was an inconclusive result [61]. The Austrian case involved correlation of dog hairs found on a murder victim to dog hairs in a suspect's car [62]. The hairs from the victim did not match any of the hairs found in relation to the suspect [62]. In Portugal, it was alleged that a young girl had been murdered and thrown into a pig sty, where small bone fragments were found [63]. After sequencing of the mtDNA, it was concluded that the bone fragments were those of pig (*Sus scrofa*) [63]. In a second case, intestines were found in a forest and thought to be of human origin but did not produce a result with human STR tests [63]. Sequencing of the cytochrome *b* gene was performed and it was concluded that the sample originated from a pig (*Sus scrofa*) [63].

The same principle has been used to identify protected primate species sold as bushmeat, turtle products [64], snake skins [65], and whale meats [66]. The benefit of the test is that even from cooked meats or skins molded into handbags, there is sufficient mitochondrial DNA present to obtain a section of the cytochrome *b* gene. So long as there is a reference sequence in the EMBL-Bank or GenBank, a comparison is possible.

4.8.1 Case Example

An example of extreme sensitivity is that of the test for the presence of the Tibetan antelope (*Pantholops hodgsonii*), otherwise called the chiru [67]. Hairs from this animal are woven into a shawl called a Shahtoosh. This antelope inhabits the high-altitude plains of Tibet and China and produces very fine hair to protect the animal from the environment. The hair from the animal is woven to make shawls of the finest quality that are highly sought after, but it takes between three and five dead antelope to make one shawl. This has led to the catastrophic decline in the numbers of the species, and it is now listed in Appendix I of CITES. Microscopic examination of the hairs taken from the shawls is possible. Hair comparison (chapter 3) plays a crucial role in species identification, although hairs from very closely related species look very similar and it requires skill, and the subjective judgement of the examiner,

to be able to determine if a hair from a shawl is a conclusive match to that of the Tibetan antelope. The hairs contain trace amounts of DNA, even after the weaving process in creating the shawl, and these hairs can be recovered and the trace amounts of DNA can be extracted. Fine hairs can be recovered and any cells adhering to the outer surface, such as from human contact, can be removed. Mitochondrial DNA will be present at exceptionally low levels; however, recent advances have permitted the development of DNA-based tests to identify the presence of the Tibetan antelope from hairs taken from shawls. The mitochondrial DNA within the shawl can be isolated and primers used to amplify a small section of the cytochrome *b* gene specific to antelope (see figure 4.9). Under standard conditions even this PCR could not amplify sufficient DNA from the hairs to be detected. A second primer, inside the first, was used to amplify from the PCR products, and a 311 bp fragment, as expected, was produced. This technique, called nested PCR, increases the sensitivity of the test and, if the primers are designed carefully, increases the specificity. The problem with using nested PCR is that with the increase in sensitivity there is a corresponding increase in the chance of contamination of the sample. It is necessary that negative controls are performed and that there is a physical separation of the laboratory where questioned samples are analyzed compared to the positive control voucher specimens.

The 311 base pair fragment was sequenced and found to match that registered in the EMBL-Bank database with 100% homology to *Pantholops hodgsonii*. Cashmere is the nearest material that could be confused with Shahtoosh, but of the 311 bases identified there are 38 differences between the Tibetan antelope and the source of cashmere (*Capra hircus*).

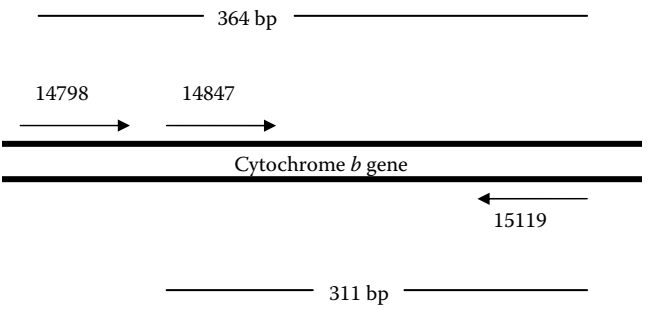

Figure 4.9 The size of PCR products produced by the primers designed to the Tibetan antelope. The outer primers produce a 364 bp fragment that was not detected after the standard conditions. The inner primer amplifies from the first PCR product and is specific to the Tibetan antelope producing a 311 bp fragment. The numbers represent the outermost base number of the primer used based on the human mitochondrial DNA sequence.

4.9 Other Mitochondrial Gene Loci

There are other regions of the mitochondrial DNA genome that have been used in species testing. These include the DNA sequences that encode ribosomal RNA molecules, particularly the 12S rRNA gene. The same methodology is used to examine DNA sequence variation as for the cytochrome *b* gene. As the DNA sequences are highly conserved, it is possible to use universal primers that will amplify a section of the gene. This portion of the gene is sequenced and the DNA sequences are aligned to those registered with the EMBL-Bank or GenBank.

The reason for using 12S rRNA rather than the cytochrome *b* gene is in the case of insufficient sequence variation at the cytochrome *b* gene. Although smaller in size, there are cases where the rRNA gene sequence shows greater variation than the cytochrome *b* gene.

This method of species identification is robust and well validated. It has been used in numerous cases to produce results that are accepted by courts throughout the world. The only issue arises when there is less than 100% homology between the questioned sample and that registered with one of the DNA databases. Due to intraspecies variation, a difference of one base over 400 DNA bases may be acceptable; however, looking at table 4.3, it should be noted that there are only four bases different between dog and wolf over the entire 1,140 bases of the cytochrome *b* gene. Clearly confidence increases with the size of the DNA section that is sequenced.

4.9.1 Bar Code for Life

In an international effort to standardize DNA loci used in species identification, an organization called the Barcode for Life Data System (BOLD) has been established [68]. A section of the cytochrome c oxidase 1 gene (COI) on the mitochondrial genome was proposed. A 648 bp section was considered to have sufficient sequence information as the DNA sequences within this section and can distinguish between 95% of the current species tested. The COI gene is a further locus in the tool of species identification.

A requirement of BOLD is that all samples registered must come from a sample whose species origin can be confirmed if required and that the laboratory providing the sequence DNA works to laboratory standards.

There is much investment in BOLD and a growing repository of DNA in the database based on part of the COI gene. Undoubtedly there will be much value in the use of BOLD as a means of species identification. Ultimately BOLD is a sequence-based species identification tool. The means of species identification described in this chapter relating to cytochrome *b* also apply to other gene loci within the mitochondrial genome.

4.10 Single Nucleotide Polymorphisms

There are many DNA bases that are highly conserved within the cytochrome *b* gene and therefore, although deciphering the entire DNA sequence is possible, there may be only a few signature DNA sequences that need to be examined. When two species are closely related, such as dog and wolf, there may only be four bases that have any significance. It is possible to develop DNA tests that interrogate these individual DNA bases only.

When there is a difference of only one base, termed a single nucleotide polymorphism (SNP), there are a number of methods to detect the DNA base. Whatever method is used, the result looks very different from the sequence data shown previously.

A small part of the DNA alignment of two closely related species, the European beaver (*Castor fiber*) and the Canadian beaver (*Castor canadensis*), is shown in figure 4.10. As can be seen in this figure, the two species share most of the DNA sequence shown.

By using a combination of primers, one of which will bind to both species and the other which will bind to either one species or the other, a species-specific PCR product is obtained. In this case, if DNA from the European beaver is present, a product of 164 bases will be produced, but if Canadian beaver is present, then the product will be 221 bases.

```
            1                                                          60
C. Beaver   ACAGCATTTTCTTCAGTCGCACACATCTGCCGAGACGTAAACTACGGCTGAATCATCCGG
E. Beaver   ACAGCATTTTCTTCAGTCGCACACATCTGCCGAGACGTGAACTACGGCTGAATCATCCGG
            ************************************ **********************

            61                                                         120
C. Beaver   TACCTTCACGCCAACGGAGCATCCATATTCTTCATCTGCCTGTTCCTACACGCAGGACGA
E. Beaver   TACCTTCACGCCAACGGAGCATCCATATTCTTCATCTGCCTGTTCCTACACGCAGGACGA
            ************************************************************

            121                                                        180
C. Beaver   GGGGTCTACTATGGATCATATACCTTCACAGAGACATGGAACATTGGCATCATTCTACTA
E. Beaver   GGAATCTACTACGGATCCTACACCTTTACAGAAACATGAAATATTGGTATCATCTTACTA
            **  ******* ***** ** ***** ***** ***** ** ***** *****  *****

            181                                                        240
C. Beaver   TTCACAGTCATAGCAACCGCCTTTATAGGATACGTCCTCCCGTGAGGGCAAATATCCTTT
E. Beaver   CTTACAGTTATAGCCACTGCTTTCATAGGATATGTACTCCCATGAGGACAAATATCCTTC
             * ***** ***** ** ** ** ******** ** ***** ***** ***********
```

Figure 4.10 Alignment of part of the cytochrome *b* gene for the European beaver and the Canadian beaver. The bases underlined at bases 1–22 act as a PCR primer site for both species. A second primer is made to the European beaver sequence at the base positions 138–164 and is specific for that species and not the Canadian beaver, whereas a primer made to the Canadian beaver sequence at bases 194–221 will bind to the Canadian beaver DNA and not the DNA from the European beaver.

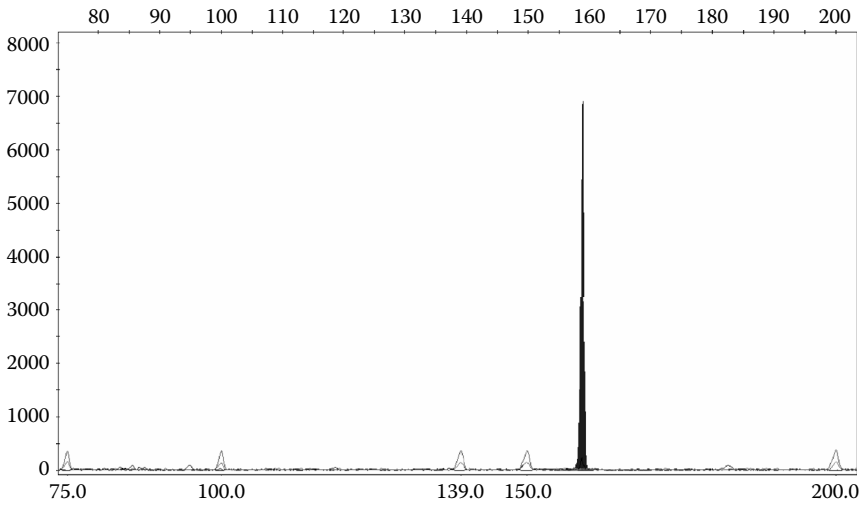

Figure 4.11 A DNA product as the expected position if the unknown was a European beaver. No DNA product was obtained at the position expected for the Canadian beaver.

Using the test shown in figure 4.10, an unknown sample suspected of being beaver can be tested. If beaver is present, only one of the two expected products should be produced. An example of the result is shown in figure 4.11.

Deciphering the full DNA sequence from a section of the mitochondrial genome, such as the cytochrome *b* gene examples, gains the most sequence information. The example of two closely related species shown in the beaver comparison illustrates that within a DNA sequence there may be only a few DNA bases that are informative in separating the two species. These may be a small number of SNPs dotted along the DNA sequence. Using SNP sites can differentiate closely related species, although the amount of sequence information is limited by the number of SNPs detected. In the case above, only one SNP is shown, although more than one SNP site can be examined at any one time. If a number of SNPs for each species are incorporated into the test, then the confidence in the identification of the species will also increase.

4.11 Mixtures

A benefit of using SNP loci is that, unlike DNA sequence information, mixtures of two or more species can be identified. Mixtures of species, particularly when human is one of the species, occur regularly, preventing conventional DNA sequencing generating interpretable results. The use of DNA sequences on the mitochondrial genome permits analysis of DNA from a fraction of one cell. This benefit comes with the problem of inevitable contamination of the

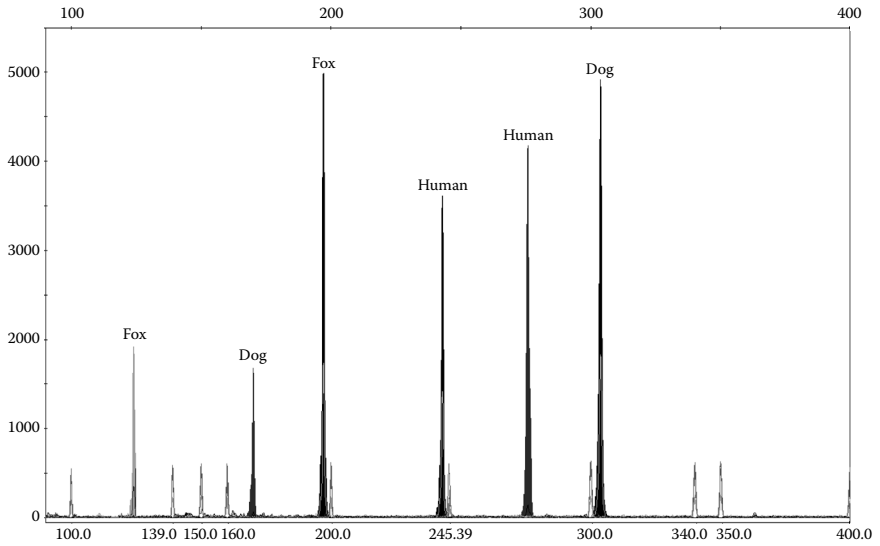

Figure 4.12 A number of colored peaks corresponding to the presence of dog (*Canus lupus familiaris*), European red fox (*Vules vulpes*), and human. The analysis can lead to a prosecution for the illegal killing of a fox with dogs.

sample with traces of human DNA. Cases investigated in the forensic science laboratory include the examination of unknown bloodstains, such as stains thought to be from an animal that has been killed illegally. The removal of a stain from clothing worn next to the skin will result in both the DNA from the white blood cells and skin cells from the wearer of the clothing.

In the investigation of the illegal killing of a fox by dogs, a DNA-based test that can identify the presence of fox and possible contaminating DNA from other animals, such as the dog, was developed. In a way similar to the beaver test, DNA primers were designed to a range of mammalian species including the European fox, the domestic dog, and human DNA sequences [69]. Human primer sets are included to detect the presence of trace levels of human DNA from persons who may have handled the item. The results of the test are shown in figure 4.12, where there is a result for all three species.

It is possible to extend these types of SNP tests further and have the ability to examine numerous species in the one test [69]. Most samples examined in the laboratory will only contain one species plus a human contaminant, but as the nonhuman species is unknown, a test can be used that will produce a characteristic result if at least one of the species is present.

The unlawful killing of animals may be for a number of reasons. This includes poaching for meat such as bushmeat; killing for sport such as bear species; killing for a product such as mementos, clothing including shoes, and bags; and killing for perfumes and supposed medicines. The range of products allows for identification either by microscopy, if possible, or by DNA.

Figure 4.13 An example of a traditional East Asian medicine product. Inside the wrapper were six sheets of plaster as used to apply to the skin. On the reverse of the outer wrapper, written in Mandarin, were the products including tiger, leopard, and musk deer. If any of these species were present, then, firstly, the importation breaks CITES regulations merely by stating that they contain listed species and, secondly, to prove that CITES-listed species are present will require a highly sensitive test.

DNA typing from cooked meat, such as goose products in salami or porpoise from steaks, has been used with much success based on the standard DNA sequencing of part of the mitochondrial DNA.

The identification of rhino from bone and sculptures has been performed as a further example of how the sensitivity and specificity of DNA typing can be applied to wildlife crimes [70].

4.11.1 Case Example

The identification of tiger products in a wide range of possible samples has led to a number of DNA-based tests. An example of this type of testing is in the examination of traditional East Asian medicines, where products are sold as containing tiger products; an example is shown in figure 4.13. In traditional East Asian medicines products there may be a range of animal species, as advertised on the product shown in figure 4.13. The species advertised as present also included leopard, although not stipulated which species of leopard, and musk deer.

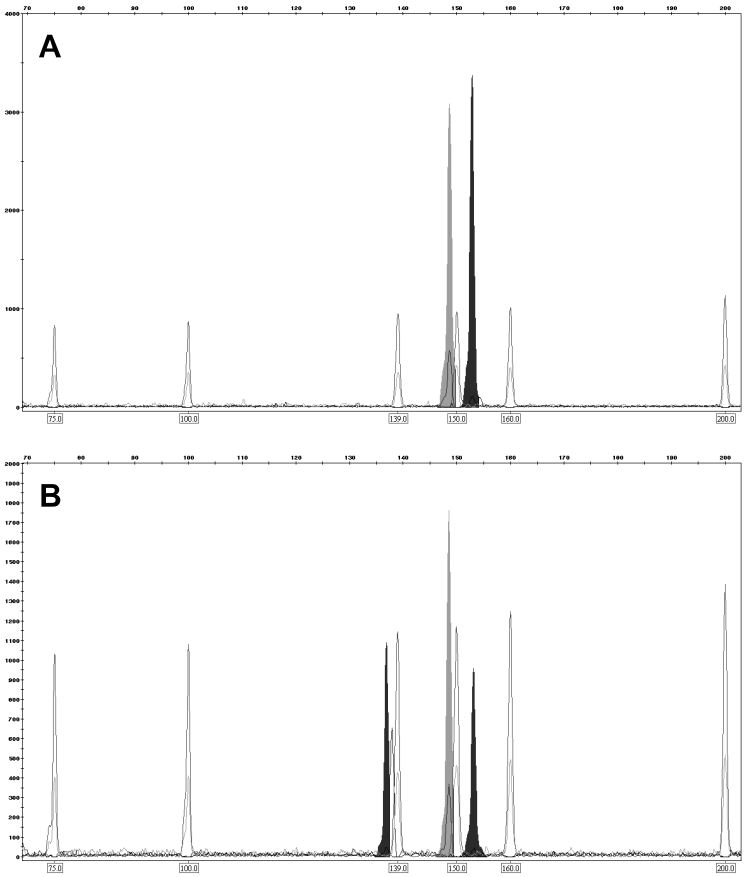

Figure 4.14 Two different samples claiming to contain tiger. (A) Two DNA products corresponding to the expected result if any member of the *Panthera* genus were present in the traditional East Asian medicine. (B) An additional DNA product is present that is produced if a member of the tiger species were present.

A test was developed that would identify all members of the *Panthera* genus, which includes all the tiger species and all the leopard species. This was accomplished by designing DNA primers that would amplify a section of the cytochrome *b* gene only if *Panthera* species are present. A second test was used to amplify a section of the cytochrome *b* gene specific to tiger species. The result of the test is shown in figure 4.14.

The results shown in figure 4.14 indicated that tiger was present in one of the traditional East Asian medicine samples provided and that the other contained a member of the genus *Panthera*. These samples were obtained in Europe and imported from China, making their importation contrary to CITES regulations.

4.12 Problems with Mitochondrial DNA Markers

Part of the benefit of using DNA loci on the mitochondrial DNA is that they are inherited down the maternal line. The benefit is that all members of the population with a common maternal ancestor will have a similar DNA type. The problem comes when hybrid species are produced by the mating of two closely related species, one of which is protected and the other which is not. If the mother is from the nonprotected species, all the offspring will have the same mitochondrial DNA type with no mixture from the protected male species. This problem has occurred within species of sturgeon that produce caviar, some of which are on CITES I. Other closely related species that produce inferior caviar are able to breed with this species but are not offered the same protection. A similar issue arises with many deer species, where the hybrid may not be subject to the legal protection of the original species. In such cases there are DNA markers in the nuclear genome that can be used, although to date less attention has been placed on these loci.

4.13 The Future of Species Testing

There are an increasing number of DNA sequences from a range of different species being lodged with the DNA database. Organizations such as BOLD aid in the standardization of loci used in standard sequence-based methods of species identification. The majority of samples tested will be from a single species, and hence the sequence comparison methods described in this chapter will remain the standard methods of species identification. For those samples where there is a mixture, the method of species identification will change to SNP testing as described in section 4.9.

References

1. Schulz, I., et al., Examination of postmortem animal interference to human remains using cross-species multiplex PCR. *Forensic Science, Medicine, and Pathology*, 2006, 2(2): 95–101.
2. Haglund, W. D., Rodents and human remains, in *Forensic Taphonomy: The Postmortem Fate of Human Remains*, W. D. Haglund and M. H. Sorg, Editors, 1997, CRC Press: London, pp. 405–414.
3. Haglund, W. D., Dogs and coyotes: Postmortem involvement with human remains, in *Forensic Taphonomy: The Postmortem Fate of Human Remains*, W. D. Haglund and M. H. Sorg, Editors, 1997, CRC Press: London, pp. 367–381.
4. Bartlett, S. E. and W. S. Davidson, FINS (Forensically Informative Nucleotide Sequencing)—A procedure for identifying the animal origin of biological specimens. *BioTechniques*, 1992, 12(3): 408–411.

5. Hillier, M. L. and L. S. Bell, Differentiating human bone from animal bone: A review of histological methods. *Journal of Forensic Sciences*, 2007, 52(2): 249–263.
6. Casper, R. M., et al., Detecting prey from DNA in predator scats: A comparison with morphological analysis, using Arctocephalus seals fed a known diet. *Journal of Experimental Marine Biology and Ecology*, 2007, 347(1–2): 144–154.
7. Butler, R., What is a species? *Chemistry and Industry*, 2003, 11: 7.
8. Mallet, J., A species definition for the modern synthesis. *Trends in Ecology & Evolution*, 1995, 10(7): 294–299.
9. Voet, D., J. Voet, and C. Pratt, Life, in *Fundamentals of Biochemistry*, Upgrade Edition, 1999, John Wiley & Sons, Inc.: Brisbane, Chapter 1, pp. 3–21.
10. Mallet, J., Species concepts, in *Evolutionary Genetics: Concepts and Case Studies*, C. Fox and J. Wolf, Editors, 2006, Oxford University Press: New York, pp. 367–373.
11. Olsen, S. J., The fossil ancestry of Canis, in *Origins of the Domestic Dog*, 1985, The University of Arizona Press, Chapter 1, pp. 1–14.
12. Verma, S. K., et al., Was elusive carnivore a panther? DNA typing of faeces reveals the mystery. *Forensic Science International*, 2003, 137(1): 16–20.
13. Hsieh, H.-M., et al., Cytochrome *b* gene for species identification of the conservation animals. *Forensic Science International*, 2001, 122(1): 7–18.
14. Saferstein, R., Identification and grouping of bloodstains, in *Forensic Science Handbook*, 1982, Prentice Hall, Inc.: New Jersey, Chapter 7, pp. 267–296.
15. Balitzki-Korte, B., et al., Species identification by means of pyrosequencing the mitochondrial 12S rRNA gene. *International Journal of Legal Medicine*, 2005, 119(5): 291–294.
16. Prakash, P. S., et al., Mitochondrial 12S rRNA sequence analysis in wildlife forensics. *Current Science*, 2000, 78(10): 1239–1241.
17. Macedo-Silva, A., et al., Hamburger meat identification by dot-ELISA. *Meat Science*, 2000, 56(2): 189–192.
18. Parson, W., et al., Species identification by means of the cytochrome *b* gene. *International Journal of Legal Medicine*, 2000, 114(1): 23–28.
19. Olsen, S. J., Prehistoric dogs in Europe and the Near East, in *Origins of the Domestic Dog*, 1985, The University of Arizona Press, Chapter 6, pp. 71–78.
20. Bottero, M. T., et al., A multiplex polymerase chain reaction for the identification of cows', goats' and sheep's milk in dairy products. *International Dairy Journal*, 2003, 13(4): 277–282.
21. Bravi, C. M., et al., A simple method for domestic animal identification in Argentina using PCR-RFLP analysis of cytochrome *b* gene. *Legal Medicine*, 2004, 6(4): 246–251.
22. Burton, R. S., Molecular tools in marine ecology. *Journal of Experimental Marine Biology and Ecology*, 1996, 200(1–2): 85–101.
23. de los Angeles Barriga-Sosa, I., et al., Inter-specific variation of the mitochondrial r16S gene among silversides, "Peces Blancos" (Atherinopsidae: Menidiinae) and its utilization for species identification. *Aquaculture*, 2005, 250(3–4): 637–651.
24. Guha, S. and V. K. Kashyap, Molecular identification of lizard by RAPD & FINS of mitochondrial 16S rRNA gene. *Legal Medicine*, 2006, 8(1): 5–10.

25. Ishizaki, S., et al., Molecular identification of pufferfish species using PCR amplification and restriction analysis of a segment of the 16S rRNA gene. *Genomics and Proteomics*, 2005, 1(1): 139–144.
26. Partis, L., et al., Evaluation of a DNA fingerprinting method for determining the species origin of meats. *Meat Science*, 2000, 54(4): 369–376.
27. Partis, L. and R. J. Wells, Identification of fish species using random amplified polymorphic DNA (RAPD). *Molecular and Cellular Probes*, 1996, 10(6): 435–441.
28. Sasazaki, S., et al., Development of breed identification markers derived from AFLP in beef cattle. *Meat Science*, 2004, 67(2): 275–280.
29. Sunnucks, P., Efficient genetic markers for population biology. *Trends in Ecology & Evolution*, 2000, 15(5): 199–203.
30. Verkaar, E. L. C., et al., Differentiation of cattle species in beef by PCR-RFLP of mitochondrial and satellite DNA. *Meat Science*, 2002, 60(4): 365–369.
31. Woolfe, M. and S. Primrose, Food forensics: Using DNA technology to combat misdescription and fraud. *Trends in Biotechnology*, 2004, 22(5): 222–226.
32. Shankaranarayanan, P., et al., Genetic variation in Asiatic lions and Indian tigers. *Electrophoresis*, 1997, 18(9): 1693–1700.
33. Lee, J.C.-I. and J.-G. Chang, Random amplified polymorphic DNA polymerase chain reaction (RAPD PCR) fingerprints in forensic species identification. *Forensic Science International*, 1994, 67(2): 103–107.
34. Calvo, J. H., R. Osta, and P. Zaragoza, Species-specific amplification for detection of bovine, ovine and caprine cheese. *Milchwissenschaft*, 2002, 57(8): 444–446.
35. Black, W.C., PCR with arbitrary primers: Approach with care. *Insect Molecular Biology*, 1993, 2(1): 1–6.
36. Welsh, J. and M. McClelland, Fingerprinting genomes using PCR with arbitrary primers. *Nucleic Acids Research*, 1990, 18(24): 7213–7218.
37. Zehner, R., S. Zimmermann, and D. Mebs, RFLP and sequence analysis of the cytochrome b gene of selected animals and man: methodology and forensic application. *International Journal of Legal Medicine*, 1998, 111(6): 323–327.
38. Wolf, C., J. Rentsch, and P. Hubner, PCR-RFLP analysis of mitochondrial DNA: A reliable method for species identification. *Journal of Agricultural and Food Chemistry*, 1999, 47(4): 1350–1355.
39. Meyer, R., et al., Polymerase chain reaction–restriction fragment length polymorphism analysis: A simple method for species identification in food. *Journal of AOAC International*, 1995, 78(6): 1542–1551.
40. Ekrem, T., E. Willassen, and E. Stur, A comprehensive DNA sequence library is essential for identification with DNA barcodes. *Molecular Phylogenetics and Evolution*, 2007, 43(2): 530–542.
41. Vences, M., et al., Comparative performance of the 16S rRNA gene in DNA barcoding of amphibians. *Frontiers in Zoology*, 2005, 2(5).
42. Gelinas, B., Raising the bar on DNA coding: U of G's Biodiversity Institute will perform high-volume "bar-coding," *Daily Mercury*, 2007, Guelph, Ontario.
43. Folmer, O., et al., DNA primers for amplification of mitochondrial cytochrome c oxidase subunit I from diverse metazoan invertebrates. *Molecular Marine Biology and Biotechnology*, 1994, 3: 294–299.
44. Pereira, F., et al., Analysis of inter-specific mitochondrial DNA diversity for accurate species identification. *International Congress Series*, 2006, 1288: 103–105.

45. Kitano, T., et al., Two universal primer sets for species identification among vertebrates. *International Journal of Legal Medicine*, 2007, 121(5): 423–427.
46. Alberts, B., et al., Introduction to cells, in *Essential Cell Biology: An Introduction to the Molecular Biology of the Cell*, 1998, Garland Publishing, Inc.: London, Chapter 1, pp. 1–36.
47. Satoh, M. and T. Kuroiwa, Organization of multiple nucleoids and DNA molecules in mitochondria of a human cell. *Experimental Cell Research*, 1991, 196(1): 137–140.
48. Gray, M. W., Origin and evolution of mitochondrial DNA. *Annual Review of Cell Biology*, 1989, 5: 25–50.
49. Saferstein, R., DNA: A new forensic science tool, in *Criminalistics: An Introduction to Forensic Science*, 2004, Prentice Hall International (UK) Limited: London, Chapter 13, pp. 353–394.
50. Zhang, D.-X. and G. M. Hewitt, Nuclear integrations: Challenges for mitochondrial DNA markers. *Trends in Ecology & Evolution*, 1996, 11(6): 247–251.
51. Wolstenholme, D. R., Animal mitochondrial DNA: Structure and evolution. *International Review of Cytology*, 1992, 141: 173–216.
52. Zouros, E., et al., An unusual type of mitochondrial DNA inheritance in the blue mussel Mytilus. *Proceedings of the National Academy of Sciences of the United States of America*, 1994, 91(16): 7463–7467.
53. Butler, J. M., Mitochondrial DNA analysis, in *Forensic DNA Typing: Biology, Technology and Genetics of STR Markers*, 2005, Elsevier Academic Press: London, Chapter 10, pp. 241–298.
54. Anderson, S., et al., Sequence and organization of the human mitochondrial genome. *Nature*, 1981, 290(5806): 457–465.
55. Linacre, A., Application of mitochondrial DNA technologies in wildlife investigations—species identification. *Forensic Science Review*, 2006, 18(1): 1–8.
56. Kuwayama, R. and T. Ozawa, Phylogenetic relationships among European Red Deer, Wapiti, and Sika Deer inferred from mitochondrial DNA sequences. *Molecular Phylogenetics and Evolution*, 2000, 15(1): 115–123.
57. Ludt, C.J., et al., Mitochondrial DNA phylogeography of red deer (*Cervus elaphus*). *Molecular Phylogenetics and Evolution*, 2004, 31(3): 1064–1083.
58. Letunic, I. and P. Bork, Interactive Tree Of Life (iTOL): An online tool for phylogenetic tree display and annotation. *Bioinformatics*, 2007, 23(1): 127–128.
59. Pääbo, S. et al., Mitochondrial DNA sequences from a 7000 year old brain. *Nucleic Acid Research*, 1988, 16(20): 9775–9787.
60. Menotti-Raymond, M., V. David, and S. O'Brien, Pet cat hair implicates murder suspect. *Nature*, 1997, 386(6627): 774.
61. Savolainen, P. and J. Lundeberg, Forensic evidence based on mtDNA from dog and wolf hairs. *Journal of Forensic Sciences*, 1999, 44(1): 77–81.
62. Nussbaumer, C. and I. Korschineck, Non-human mtDNA helps to exculpate a suspect in a homicide case. *International Congress Series*, 2006, 1288: 136–138.
63. Caine, L., et al., Species identification by cytochrome b gene: Casework samples. *International Congress Series*, 2006, 1288: 145–147.
64. Hsieh, H.-M., et al., Species identification of *Kachuga tecta* using the cytochrome b gene. *Journal of Forensic Sciences*, 2006, 51(1): 52–56.

65. Wong, K.-L., et al., Application of cytochrome b DNA sequences for the authentication of endangered snake species. *Forensic Science International*, 2004, 139(1): 49–55.
66. Grohmann, L., et al., Whale meat from protected species is still being sold in Japanese markets. *Naturwissenschaften*, 1999, 86(7): 350–351.
67. Lee, J.C.-I., et al., DNA profiling of Shahtoosh. *Electrophoresis*, 2006, 27(17): 3359–3362.
68. Hebert, P. D. N., et al., Biological identifications through DNA barcodes. *Proceedings of the Royal Society B: Biological Sciences*, 2003, 270(1512): 313–321.
69. Tobe, S. S. and A. M. T. Linacre, A multiplex assay to identify 18 European mammal species from mixtures using the mitochondrial cytochrome b gene. *Electrophoresis*, 2008, 29(2): 340–347.
70. Hsieh, H.-M., et al., Species identification of rhinoceros horns using the cytochrome b gene. *Forensic Science International*, 2003, 136(1–3): 1–11.