Genetic Analyses for phylogenetic, taxonomic and identification purposes

Sylvain Eichhorn, Sonja Nadalon

Gymnasium Kirschgarten, 4dbZ

Teacher: Patrick Ruggle

27/4/2016

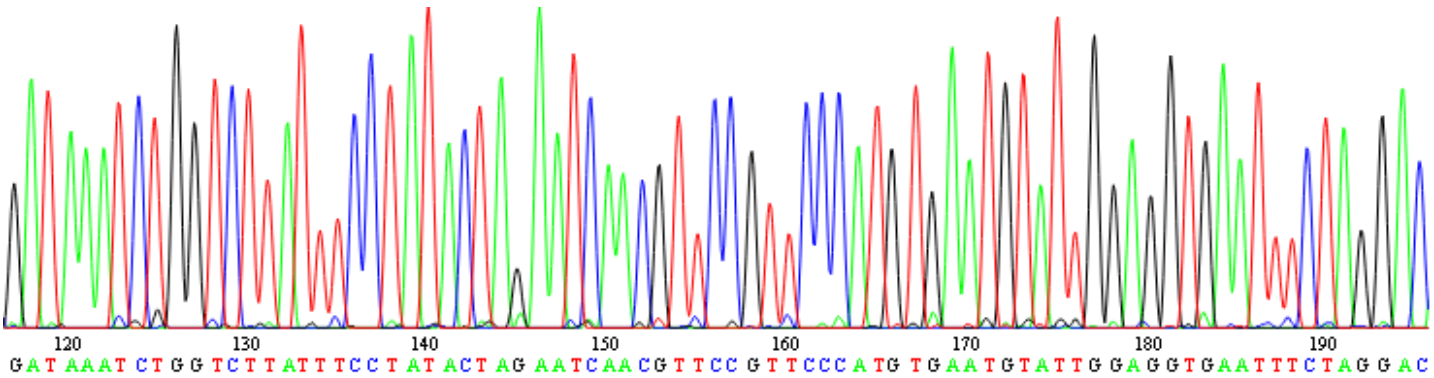

Table of Contents

1.0 Preface

Our interest in ornithology and the variety of bird species is the main reason for the choice of writing a term paper about the use of genetic analyses for taxonomic, phylogenetic and identification purposes. As a birdwatcher you are frequently confronted with difficulties in optical or acoustical species identification and furthermore with the problematic of the definition of species. You rapidly realise that the concept of species is a man-made attempt to bring an order in a complex natural system. In practise it is often difficult to make out the difference between subspecies and species as there is a fluent passage and not a distinct boundary.

With the possibilities of analysing base sequence of the DNA taxonomists obtained a new instrument for classification purpose. The degree of relationship between species can be determined by looking at specific marker in their genomes. Moreover it is possible to identify species by looking at fragments of their DNA. This also gives the opportunity of discovering new species. The actuality of this topic and the possibility of analysing evolutionary relationship between bird species makes it especially interesting.

In this paper we intend to describe some methods based on genetic analyses for identification and phylogenetic as well as taxonomic purposes. Moreover we want to investigate whether the use of genetic analyses is a useful tool for classifying and identifying organisms and to derive patterns of relationship between species. We are also interested whether the use of genetic analyses helps to establish a better way of defining a species.

2.0 Introduction

Classifying the organism of the planet in a logical way has always been a challenging task for taxonomists. By looking at different features of morphology, behavioural characters and other specific characteristics the organisms were grouped in taxa in a hierarchical way. Recently the use of DNA analysis has become a current method to reveal unknown patterns of relationships between species.

In the 1950s some scientists were experimenting with techniques using molecular characteristics to classify organisms. By the 1970s Charles Sibley was doing pioneer work on DNA-DNA hybridization, a method to determine the degree of relationship between species. His works revealed that the systematic of birds at that time needed a significant revision. The invention of Sanger sequencing in 1977 made it possible to determine the nucleotide sequence of genes. By comparing sequences of different species it is possible to compile phylogenetic trees, which shows the probable evolutionary history and relationship of species. From then on the methods were steadily improved and became more efficient (AUTHOR UNKNOWN 1 2016).

Nowadays genetic analyses are also being used to identify species by looking at a particular gene region. This method known as DNA-Barcoding was proposed by Paul Herbert of the University of Guelph in Ontario, Canada, in 2003 (CBOL 2016).

3.0 Description of the methods

3.1 DNA-DNA Hybridization

DNA-DNA Hybridization is a quite simple method to analyse how closely two species are related to each other. The DNA of two species is required. The DNA-double helices are denatured through heat supply, which results in single strands of DNA. The single strands of DNA of the two different species are then allowed to form hybrid DNA double helices. The closer related two species are, the more should their DNA have similar base sequences. The energy required to separate the strands of closely related species is therefore higher than for less related species, because more bases of the strands match by complementary base pairing, forming more hydrogen bonds, which hold the double helices together. The heat energy needed to separate the hybrid double helix therefore lowers if the genetic distance between two species increases. The hybrid DNA is exposed to heat and the temperature required to separate its strands is compared to the temperature needed to separate the strands of the double helix of a single species. This enables you to make a statement if species are closely related or not (THE GALE GROUP INC. 2003).

3.2 Derivation of relationship between species using genetic marker

A more precise method to analyse the evolutionary relationship of species uses so called markers, which consist of a specific region in a gene. The markers get amplified by PCR and the nucleotide sequence of the markers is determined by DNA sequencing and then compared to markers of other species. It is then possible to create a phylogenetic tree by comparing the genetic distance between species.

3.3 DNA sequencing

The sequencing of genes is of major importance in modern taxonomy and phylogenetics. In DNA sequencing the order of the four different nucleotides, known as the nucleotide sequence, is determined in markers. This is a crucial procedure to be able to compare markers and to derive evolutionary relationship between different species. There are different approaches to DNA sequencing. Sanger sequencing was the most widely used sequencing method for 25 years. For this method four separated but equivalent reactions are done, one for each type of base. In each reaction templates of DNA are duplicated by adding Primers, which bind at the beginning of the markers, DNA-polymerase, the four different nucleotides and one di-deoxynucleosidetriphosphates (dNTP) containing one of the four bases. The dNTP is a kind of modified nucleotide,

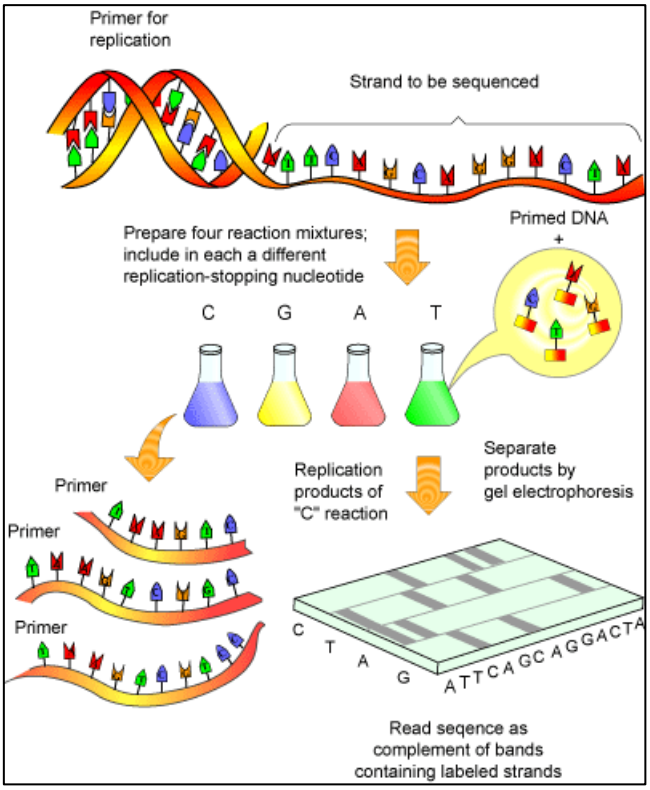

Fig. 1: Sanger sequencing (BIOLOGY MULTIPLE CHOICE QUIZZES 2015)

which lacks a hydroxyl group at the 3' end and cannot form a phosphodiester bond, causing the extension of the new strand to cease when it gets implemented. The dNTP is added in a 100-fold lower concentration than the nucleotide containing the same base, so that the new strands do not get ceased always at the beginning, but at all the different positions in the marker containing the complementary base of the dNTP. This results in strands of different length, always ending at a position with the base in the dNTP. For each of the four reactions a dNTP with another of the four bases is used, so that you finally get strands ending at every position of the gene. The synthesised strands of every reaction are separated by their length by gen electrophoresis and arranges in a way that it is possible to determine the positions of each base. If the sequencing reactions were successful there should be one mark for each position in the locus corresponding to a specific base on the gel and it is possible to read off the nucleotide-sequence of the marker by applying the rules of complementary base pairing. Figure 1 shows the explained steps of Sanger sequencing graphically. The nucleotide sequence of the marker can be read automatically and gets depicted in a graph like the one on the front page of this paper.

New methods for the sequencing have been developed under the umbrella term next generation sequencing, which are faster and can sequence more bases (AUTHOR UNKNOWN 2 2016).

3.4 Sequence Alignment

If the nucleotide sequences of two or more markers are compared to derive patterns of relationship between species, they have to be aligned. In sequence alignment the positions of these sequences are compared and aligned that a maximum of identical positions are achieved and similarities in the nucleotide sequences are highlighted. Therefore you have to take similar sequences, as well as frameshift mutations caused by insertion or deletion into account. The sequences are arranged in a way that homologous positions are placed below each other. The homologous positions in different markers descend with a high probability from a common ancestor. The genetic distance between the species is determined by looking at the percentage of positions, where a mutation caused a difference in the base sequences of the species. Sequences are often aligned by computers with the help of algorithms. The aligned sequences can be used to generate a phylogenetic tree of the corresponding species. Hyphens are used to compensate a frameshift mutation caused by the insertion or deletion of one or more bases, as you can see in Figure 2 (AUTHOR UNKNOWN 3 2016).

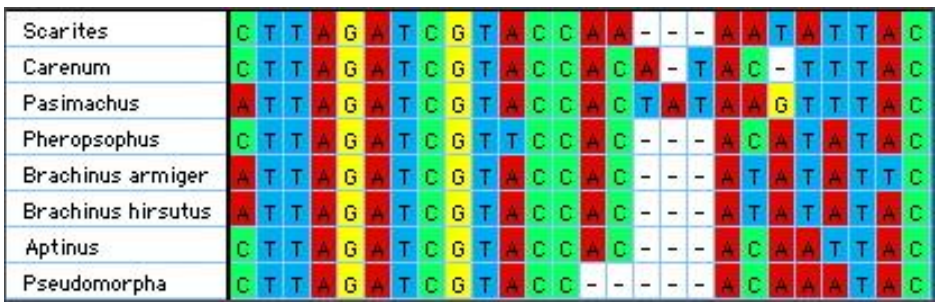

Fig. 2: Example of a sequence alignment with eight sequences (AUTHOR UNKNOWN 4 2016)

3.5 DNA-Barcoding

DNA-Barcoding is a method to identify species by looking at a short specific marker in their DNA. This marker should be very similar within one species but vary between different species. To identify an unknown specimen its DNA barcode can be compared with the available barcodes of different species in a database. The specimen then can be matched to the corresponding species. The base sequence of the DNA barcode is therefore comparable to the black lines of barcodes on products in supermarkets and the basic principle is pretty much the same (INTERNATIONAL BARCODE OF LIFE 2016).

For this method to work, the gene region used for the barcode has to fulfil some criteria. The chosen locus should vary largely between different species, included closely related ones, and be very constant within the same species. Moreover amplification and sequence alignment has to be easily possible for the marker. Evolutionary conserved primer sites should be present at the beginning and ending of the locus of interest so that the same pair of primer can be used for a large variety of different organisms to ensure an easier amplification.

For animals a 648 base-pair long locus of the mitochondrial cytochrome oxydase I gene (COI) has been recognised as the official barcode marker by the International Barcode of Life Project. This project aims to create an international library of barcodes of many species. It can be used very effectively to identify birds, butterflies, fish, flies and many other animal groups.

The COI in plants and fungi evolves too slowly and is therefore not a suitable region for the barcodes of plants as there is not enough variation between the barcodes of different species. In fungi this gene contains large introns which tend to mutate more easily. For fungi a nuclear gene is recognized as the official barcode marker, whereas two genes located on the chloroplast genome are the official barcode markers for plants (AUTHOR UNKNOWN 5 2016).

3.6 The Generation of a DNA-Barcode

For generating a DNA-Barcode a specimen is needed. It can come from various sources as collections of natural history museums, seed banks, zoos, botanical garden, and herbariums or it can be collected in the field. A small piece of tissue is removed from the specimen to extract its DNA. There are different methods for the extraction; however, the basic principle is always the same. The cells in tissues are lysed and the lipids in the membrane and proteins associated with the DNA are removed to finally precipitate and isolate the DNA. The DNA molecules can get stored in freezers or dried and preserved at room temperature. The barcode marker is replicated using PCR amplification. In this process a primer pair is used to set the starting points of the DNA synthesis on both strands of the template. The specific gene which needs to be replicated gets therefore defined. The gene is amplified so that the nucleotide sequence can then be determined. The result of the base sequence analyse reveal the barcode of the species the specimen belongs to. The finished barcode is placed in a reference database where it can be used to identify other specimen by comparing their barcodes to the ones available in the database. In addition to the barcode all the collection data concerning the specimen are entered into the database, as for example the finding location and the date. The specimen itself gets deposited in a collection as a voucher (AUTHOR UNKNOWN 6 2016).

As you can see in Figure 3, the barcode can be represented as a series of lines in four different colours, coding for the four bases cytosine, guanine, adenine and thymine.

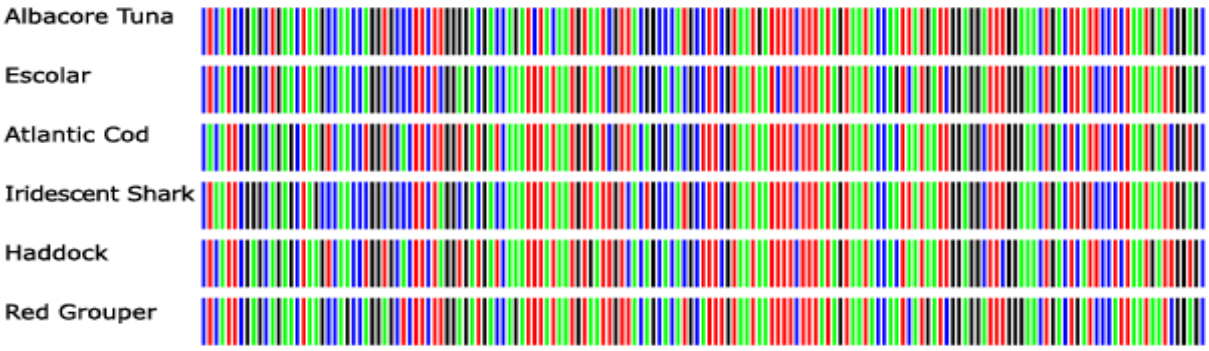

Fig. 3: Barcodes of fish species (BIODIVERSITY INSTITUTE OF ONTARIO 2014)

4.0 Documentation and pictures of research institutions visited

4.1 Interview with Dr. Manuel Schweizer

Dr. Manuel Schweizer is curator at the Museum of Natural History in Bern and responsible for its collection of birds. Research is his main occupation. He wrote a master thesis about population genetics of common vole (field mouse) in the Grosses Moos in Switzerland and a dissertation about the evolutionary diversification and biogeography of parrots for which he used a lot of molecular genetic analyses. He was one of the discoverers of the breeding grounds of the very rare Large-billed Reed Warbler in Tajikistan and made genetic analyses to confirm the identification. He is especially interested in birds and ornithology.

Which method is usually used to analyse relationship between species?

DNA is extracted from the cells of specimens and specific genes of interest, so called markers, are amplified by PCR. If the PCR was successful the markers are sent to a laboratory, where they get sequenced. Formerly Sanger sequencing was used to determine the base sequence of genes but nowadays more efficient methods are used, which enable a more effective sequencing. The sequences are aligned by a computer programme with complex algorithms. This allows comparing the base sequences of different species as the algorithms allocate homologous base pairs and substitutions are detected. The genetic distance of two species is determined. This data can be used to create a phylogenetic tree. The substitution rate of some genes is known. It is hence possible to calculate when there was a separation of a taxon and it is possible to reconstruct the history of origin of specific species.

Is DNA-DNA Hybridisation still used to derive relationship between taxa?

DNA-DNA Hybridisation is not used anymore in research since about a decade as it is not precise in comparison to more modern methods. However, several findings made with DNA-DNA Hybridisation appeared to be correct after more recent investigations with more accurate methods but others are not.

Is it possible to compile a phylogenetic tree only by means of genetic analyses?

Phylogenetic trees can be made by using only data of genetic analyses. The result can then be compared to morphological traits as these often give evidence about the relationship between species.

How many different sequences are analysed per species in studies about the relationship between species or populations?

The more sequences you analyse the better result you get. In a current project we look at five different sequences on nuclear DNA and mitochondrial DNA, this being more at the lower limit.

Are there any species which cannot be determined by DNA-Barcoding or where other problems arise?

The need of a barcoding gap is necessary to be able to identify two species doubtless. For this the intraspecific genetic variation and the interspecific genetic variation of species is compared. This means that the variation within the gene of specific species is compared with the variation between the genes of these species. If the intraspecific variation is smaller than the interspecific variation there is a barcoding gap. If this is not the case there is an overlap and some barcodes are not identifiable. The idea that DNA-barcoding will once replace identification experts is therefore not practicable. The graph on the right should help to understand this principle (figure 4)

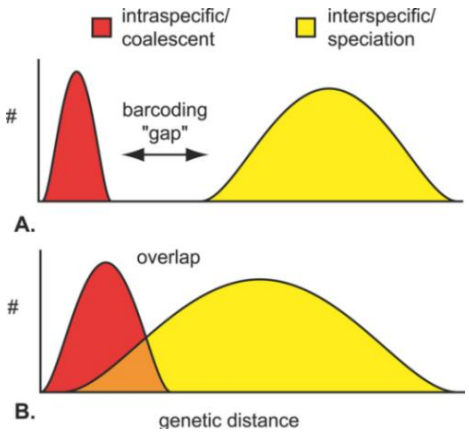

Fig. 4: Barcoding gap
(MAYER AND PAULAY 2005)

Are introns of disadvantage in genetic analyses?

In barcoding a goal is to be able to use the same primers for a wide range of organisms. If you would have introns in a barcode marker this would be more difficult as they are more variable between different groups. Therefore it is of advantage for barcoding that the mitochondrial gene of animals has no introns. However, in nuclear gene introns can be very interesting as they do not code for proteins and mutation do not affect the organism. The mutation can be passed over to the next generation more easily as selection does not ensure that it gets eliminated and the introns are more variable in different species. This can give more detailed information about the degree of relationship between species.

To what extent does phylogenetic research supports the conservation of nature?

We are doing fundamental research which is not necessarily intended for conservation projects. Of course it is favourable if the research we do can also be used in nature conservation. Some phylogenetic projects on little-known taxa reveal the existence of undescribed species which can then eventually be protected. DNA-Barcoding can also be used for biomonitoring which is a fundament of species conservation. With DNA-Barcoding it is for example possible to identify tadpoles and to allocate it with the frog and to learn more about their needs in both of their development stages.

How much does a genetic analyse costs?

The DNA extraction of a specimen cost about three Swiss francs, the chemicals needed for the PCR reaction about 50 raps. Sometimes a purification is needed for about one and a half francs. The sequencing of one sequence costs about four francs and is done abroad, so you get a sequence for less than 10 francs.

Which method was used to identify the Large-billed Reed Warbler?

The Large-billed Reed Warbler was identified by comparing the base sequence of the cytochrome b gene in the mitochondria of the caught individuals and a voucher in a museum.

The subalpine warbler was assumed to be a polytypic species with five subspecies but recently evidence has come up that there are at least two species if not three. Do a lot of genetic studies underlie this new species complex?

Genetic analyses gave occasion to reconsider the species complex of the subalpine warbler and played a decisive role in establishing a new way of classifying its different taxa. Genetic analyses and the corresponding phylogenetic tree of the former subspecies showed that three genetically distinct groups exist. It is now clear that the subalpine warbler complex includes at least two different species as it is known that they breed sympatrically without mixing. The existence of a third species is a bit a question of definition. For these analyses cytochrome b was used as marker.

Acknowledgement

We want to thank Manuel Schweizer for taking time to answer all of our questions. His statements and the explanations he gave us were very helpful and have contributed a lot to this term paper. We were also allowed to take pictures of the laboratory and he gave us a very interesting insight into latest research.

4.2 Pictures of equipment in the laboratory of the Museum of Natural History in Bern

Fig. 5: Thermal cycler

Fig. 6: Purification device

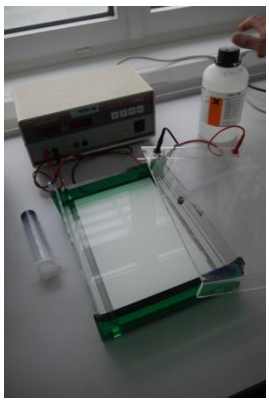

Fig. 7: Gel electrophoresis

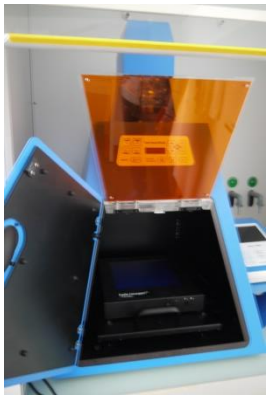

Fig. 8: UV-light chamber

Figure 5: The DNA is amplified with a thermal cycler, which is able to heat and cool down the DNA at preprogrammed intervals to initiate the different phases of the PCR amplification.

Figure 6: The amplified DNA fragments are purified automatically.

Figure 7: To verify whether the PCR worked, gel electrophoresis is used. In gel electrophoresis the amplified DNA is put on a gel with an electric field. The negatively charged DNA is attracted to the positively charged electrode. Short strands of DNA move faster through the gel particles and cover a bigger distance than longer strands.

Figure 8: After the electrophoresis, the gel is put under UV-light and the DNA gets visible. If the PCR amplification worked there should be a lot of short strands, which move at the same speed. Under the UV line you would see a clearly defined mark. If the PCR reaction did not work well there would be strands of different lengths and the mark would be diffuse and spread over a larger surface of the gel as the strands would move at different speeds.

5.0 Discussion

The introduction of genetic analyses for taxonomic and phylogenetic purposes has given new opportunities to successfully analyse patterns of relationship between taxa. Furthermore, genetic analyses of mostly little-known taxa have given evidence for the existence of several undescribed species. As genetic analyses have not been used for a very long time, there is still a lot of research, which can be done in this field.

Because the efficiency of the methods has been steadily improved, the procedure is simpler and you get more exact results nowadays. The possibility of sequencing genes has literally revolutionized the use of genetic analyses as the exact nucleotide sequences of species could now directly be compared. The introduction of next-generation sequencing for faster large-scaled sequencing and the possibility of evaluating data faster and more precisely thanks to more powerful computers has simplified investigations about evolutionary relationship of species. New methods are not comparable to the earliest methods like DNA-

DNA hybridization. It became possible to create detailed phylogenetic trees based on genetic data instead of morphological aspects for example.

The introduction of DNA-Barcoding gave a new opportunity for identifying cryptical species. Barcoding can also be used for biomonitoring projects and therefore for the protection of species. In many countries there are projects aiming to create databanks with the barcodes of as many as possible organisms existing in the country. Although the idea that barcoding will enable everybody to identify every specimen is not realistic, because there are cases, where a proper identification of two closely related species is only possible with the help of a specialist, this would be of great value to identify a large range species. It also facilitates the identification of larval, or other developmental stages of species, which would otherwise be unidentifiable and it can even be used to assign the males and females of dimorphic species as belonging to one species (AUTHOR UNKNOWN 7 2016). The needs of a species in its entire life cycle can therefore be studied more easily.

Genetic analyses for phylogenetic, taxonomic and identification purposes cannot be considered as being a highly controversial topic like other application of genetics for example in agriculture or medicine. Some people are astonished that vouchers are still collected in times of digital photography. However, if the specimen turns out to be an undescribed species the morphological aspects will have to be studied in detail and the specimen can be used to perform further genetic analyses. Only if the species gets described, steps for its protection can be initiated, so for the population as a whole it is beneficial.

Concluding, it is obvious that genetic analyses are a useful tool for the derivation of evolutionary relationships and for the identification of many species, where an optical or acoustical identification would cause problems. The new classification of the taxa of the Subalpine warbler complex is an example, which shows that genetic analyses can clarify the systematics of some closely related species. Even if genetic analyses simplify the classification of organisms, they do not simplify the definition of a species as this remains an artificial concept, which does not fit for all the organisms of the planet ideally.

6.0 Summary

Preface

The use of genetic analyses for taxonomy, phylogeny as well as for the identification of species is an actual topic and as ornithologists we are especially interested in it. Our aim was to describe some methods used for genetic analyses and we asked ourselves whether genetic analyses are favourable for taxonomic, identification and phylogenetic purposes.

Introduction

Genetic analyses are a more recent tool for classifying organisms and deriving evolutionary relationship of species. The beginnings of molecular phylogenies were in the 1950's. The methods have been steadily improved so that genetic analyses are getting more and more precise and also simpler. The sequencing of genes allowed comparing markers of different species and determining their degree of relationship. In 2003 DNA-barcoding was introduced to identify species by looking at a specific sequence in their genes.

Description of the methods

DNA-DNA-hybridisation is a simple method to determine how closely related two species are by hybridizing their DNAs. This method is no longer used as it is not very accurate.

For modern evolutionary relationship analyses, markers of different species are sequences and then aligned to be compared. Sanger sequencing was the most widely used method to sequence genes for a long time. However, in the present day there are more effective methods, known as next generation sequencing.

DNA-Barcoding is a method to identify species by looking at one specific marker, which has to vary between different species but has to be similar within a species. If the DNA-barcode of a specimen is determined it can be compared to barcodes in databases and then be matched to the corresponding species.

Documentation and pictures of research institution visited

We made an interview with Dr. Manuel Schweizer. He is doing research about the evolutionary relationship of species and is experienced in barcoding and genetic analyses.

The interview revealed some problematics of DNA-barcoding, but also manifested the uses of genetic analyses for research and sometimes even nature conservancy. Some examples were discussed concerning the species complex of the Subalpine warbler and the identification of the Large-billed Reed Warbler. In the laboratory of the museum the markers are amplified by PCR amplification but the sequencing is done abroad.

Discussion

Genetic analyses are a valuable tool for deriving evolutionary relationship between species and they can be very useful to identify species, which would be uneasily identifiable by morphological characteristics only. They have been the key for the description of new species. Genetic analyses are a good fundament to rethink the former systematic and classification of species; however, they do not eliminate some of the problematics of the definition of a species.

7.0 References

Author unknown 1 (2016). Wikipedia. URL:https://en.wikipedia.org/wiki/History_of_molecular_evolution (Last retrieved: 18. April 2016).

Author unknown 2 (2016). Wikipedia. URL:https://en.wikipedia.org/wiki/Sanger_sequencing (Last retrieved: 19. April 2016).

Author unknown 3 (2016). Wikipedia. URL:https://en.wikipedia.org/wiki/Sequence_alignment (Last retrieved: 14. April 2016).

Author unknown 5 (2016). SwissBOL. URL:http://www.swissbol.ch/swissbol/page-31965.html (Last retrieved: 20. April 2016).

Author unknown 6 (2016). SwissBOL. URL:http://www.swissbol.ch/swissbol/page-31966_en.html (Last retrieved: 20. April 2016).

Author unknown 7 (2016). SwissBOL. URL:http://www.swissbol.ch/swissbol/page-31967.html (Last retrieved: 21. April 2016).

CBOL (2016). Barcode of Life. URL:http://www.barcodeoflife.org/content/about/what-dna-barcoding (Last retrieved: 19. April 2016).

International Barcode of Life (2016). International Barcode of Life. URL:http://www.ibol.org/about-us/what-is-dna-barcoding/ (Last retrieved: 20. April 2016).

The Gale Group Inc. (2003). Encyclopedia. URL:http://www.encyclopedia.com/topic/DNA_Hybridization.aspx (Last retrieved: 23. April 2016).

Figures

Author unknown 4 (2016). Sequence Alignment. URL:http://www.sequence-alignment.com/ (Last retrieved: 24. April 2016).

Biodiversity Institute of Ontario (2014). CCDB. URL:http://ccdb.ca/research.php (Last retrieved: 24. April 2016).

Biology Multiple Choice Quizzes (2015). QuizBiology. URL:http://www.quizbiology.com/2015/02/multiple-choice-quiz-on-dna-sequencing.html#.Vx_R0Hr2OCQ (Last retrieved: 25. April 2016).

Mayer and Paulay (2005). DNA Barcoding. URL:http://dna-barcoding.blogspot.ch/2013/02/the-seven-deadly-sins-of-dna-barcoding-7.html (Last retrieved: 26. April 2016).