Approches bioinformatiques pour l'assessment de la biodiversité

Tiayyba Riaz

▶ To cite this version:

Tiayyba Riaz. Approches bioinformatiques pour l'assessment de la biodiversité. Sciences agricoles. Université de Grenoble, 2011. Français. NNT : 2011GRENV076 . tel-00716330

HAL Id: tel-00716330
https://theses.hal.science/tel-00716330

Submitted on 10 Jul 2012

UNIVERSITÉ DE GRENOBLE

THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ DE GRENOBLE

Spécialité : **Biodiversité Ecologie Environment**

Arrêté ministériel : 7 août 2006

Présentée par

Tiayyba RIAZ

Thèse dirigée par **Eric COISSAC**

préparée au sein du **Laboratoire d'Ecologie Alpine**
dans **l'École Doctorale Chimie et Science du Vivant**

Approches bioinformatiques pour l'évaluation de la biodiversité

Thèse soutenue publiquement le **23 Novembre, 2011**,
devant le jury composé de :

M Pierre BREZELLEC
MdC HDR, UVSQ, Rapporteur
M Guy PERRIÈRE
Dr-HDR, CNRS, Rapporteur
M Pierre PETERLONGO
CR, INRIA, Examinateur
M Christian BROCHMANN
Prof, Natural History Museum Oslo, Examinateur
M Serge AUBERT
Prof, UJF, Examinateur
M Eric COISSAC
MDC-HDR, UJF, Directeur de thèse

Bioinformatics Approaches For The Assessment Of Biodiversity

Abstract

This thesis is concerned with the design and development of bioinformatics techniques facilitating the use of metabarcoding approach for measuring species diversity. Metabarcoding coupled with next generation sequencing techniques have a strong potential for multiple species identification from a single environmental sample. The real strength of metabarcoding resides in the selection of an appropriate markers chosen for a particular study and thus identification at species or higher level taxa can be achieved with carefully designed markers. Moreover next generation sequencers are producing tremendous amount of data which contains a substantial level of errors that bias biodiversity estimates. In this thesis, we addressed three major problems: evaluating the quality of a barcode region, designing new barcodes and dealing with errors present in DNA sequences.

To assess the quality of a barcode region we developed two formal quantitative measures called *barcode coverage* (B_c) and *barcode specificity* (B_s). B_c gives a measure of universality of primer pairs, and B_s deals with the ability of barcode region to discriminate between different taxa. These measures are extremely useful for ranking different barcodes and selecting the best markers.

To design new barcodes for metabarcoding applications we developed an efficient program called *ecoPrimers*. Based on the above two quality indices and with integrated taxonomic information, *ecoPrimers*[1] enables us to design primers and their corresponding barcode markers for any taxonomic level. Moreover with a large number of tunable parameters it allows us to control the properties of markers. Using efficient algorithms and implemented in C language, *ecoPrimers* is efficient enough to deal with large data bases including fully sequenced bacterial genomes.

Finally to deal with errors present in DNA sequences, we analyzed a simple set of PCR samples obtained from the diet analysis of snow leopard. By measuring correlations between different properties of errors, we observed that most of the errors were introduced during PCR amplification. In order to deal with such errors, we developed an algorithm

[1] `http://www.grenoble.prabi.fr/trac/ecoPrimers`

using graphs approach, that can differentiate true sequences from PCR induced errors. The results obtained from this algorithm showed that de-noised data gave a realistic estimate of species diversity studied in French Alpes. This algorithm is implemented in program *obiclean*.[2]

Key Words:

Taxonomy, Species Inventory, Biodiversity, Paleoecology, Diet Analysis, DNA Barcoding, Metabarcoding, Environmental Sample, Barcode Markers, Coverage, Specificity, Conserved Regions, Algorithm Complexity, Metaheuristics, Mutations, High Throughput DNA Sequencing

[2] `https://www.grenoble.prabi.fr/OBITools`

Resumé

Cette thèse s'intéresse à la conception et au développement de méthodes bioinformatiques facilitant l'utilisation de l'approche "DNA metabarcoding" pour estimer la diversité des espèces dans un environement. Le DNA metabarcoding en reprenant l'idée du code barre ADN développée par Hebert *et. al* (2003) permet grâce au séquençage haut débit l'identification taxonomique des organismes présent dans un échantillon environnemental. Trois problèmes bioinformatiques seront abordés dans ce manuscrit : l'évaluation la qualité d'un code barre dans le cadre du DNA metabarcoding, l'identification de nouvelles régions du génome utilisables comme code barre ADN et l'analyse des données de séquençage afin filtrer les erreurs et de limiter le bruit masquant le signal taxonomique.

Contrairement au "DNA barcoding" classique qui utilise des marqueurs standards, le metabarcoding doit utiliser des marqueurs qui sont souvent sélectionnés et adaptés pour chacune des études. La qualité de l'identification des taxons repose donc énormément sur ce choix. Pour évaluer la qualité d'un code barre ADN, j'ai développé deux mesures permettant d'estimer de manière objective la couverture (B_c) et la spécificité (B_s) d'un marqueur. La couverture mesure l'universalité d'une paire d'amorces et donc sa capacité à amplifier par PCR (Polymerase chain reaction) un grand nombre de taxons. La spécificité, quant à elle, mesure la capacité de la région amplifiée à discriminer entre les différents taxons. Ces mesures permettent de classer des codes barres ADN et donc de sélectionner *a priori* le meilleur pour une application donnée.

Disposant de ces deux nouveaux indices de qualité d'un code barre ADN, il devenait possible de chercher a identifier la portion d'un génome les maximisant. Pour cela, j'ai développé le logiciel *ecoPrimers*[3]. Basé sur l'optimisation de ces deux mesures, *ecoPrimers* propose à partir d'une liste de séquences exemples et d'une liste de séquences contre-exemples un ensemble de paires de d'amorces qui permettent l'amplification par PCR de codes barres ADN. *ecoPrimers* possède un grand nombre de paramètres qui permettent de contrôler les propriétés des amorces et des codes barres suggérés. Ce travail a nécessité de développement d'un algorithme efficace d'identification des mots conservés dans

[3] http://www.grenoble.prabi.fr/trac/ecoPrimers

un quorum des séquences exemples et absent des contre-exemples. Il en résulte que *ecoPrimers* est suffisamment efficace pour apprendre à partir de grand jeux de données, y compris l'ensemble des génomes bactériens entièrement séquencés.

La dernière partie de ce manuscrit résume un ensemble d'observations préliminaires que nous avons réalisées sur les erreurs introduites dans les séquences tout au long du processus allant de l'échantillonnage au fichier final contenant celles-ci. Nous avons utilisée pour cela un ensemble d'échantillons de composition taxonomique simple permettant de séparer aisément le signal du bruit. Ces échantillons sont issus de l'analyse du régime alimentaire du leopard des neiges (*Uncia uncia*). La mesure de corrélations entre différents paramètres des erreurs observées dans ces échantillons, nous laisse supposer que la plupart de celles-ci se produisent durant l'amplification par PCR. Pour détecter ces erreurs, nous avons testé un premier algorithme simple basé sur une structure de graphes dirigés acycliques. Les résultats obtenus à partir de cet algorithme ont montré que les données de-bruitée donnent une estimation réaliste de la diversité des espèces pour une série d'échantillons provenant du la vallée de Roche Noire (Alpes françaises).

Mots Clés:

Taxonomie, Liste d'espèces, Biodiversité, Paléoécologie, Analyse du Régime Alimentaire, Barcoding de l'ADN, Metabarcoding, Échantillons Environnementaux, Amorces PCR, Couverture, Spécificité, Régions Conservées, Complexité Algorithmique, Metaheuristique, Séquençage de l'ADN Haut Débit

Preface

DNA barcoding has become a fairly useable method of choice for rapid species identification in the last decade. There are two principal types of DNA barcoding, the conventional barcoding applied to single specimen identification and the *sensu lato* or metabarcoding used for the identification of multiple species from single environmental samples.

Ecological studies mostly require the determination of the list of species involved in the ecological process under study. DNA metabarcoding coupled with next generation sequencing techniques provide the opportunity to produce a large amount of data for measuring biodiversity. Despite its relationship with DNA barcoding, the particularities of the DNA metabarcoding require to develop specific methodologies for data analysis. In this context, this thesis is concerned with the development of bioinformatics techniques which can facilitate the use of DNA metabarcoding for the accurate assessment of biodiversity. Due to the specific constraints imposed by metabarcoding, selection of the appropriate markers for a given ecological study and designing new optimal barcode regions is very important. Moreover the barcode data produced by next generation sequencing techniques needs be analyzed for removing the noisy reads and it is important to understand the potential sources of noise in order to make precise and unbiased diversity estimates.

This thesis contains 5 chapters. I begin with a general overview of the species inventory concept, talking about its importance and applications. The first chapter covers two major areas: first is more biological and includes topics related to the importance of biological classification systems used for species inventory and identification, well known classification methods and a detailed introduction of DNA barcoding and DNA metabarcoding. The second part is more technical where I define some computer science terms and give an overview of string matching algorithms (which are the basis for finding conserved DNA regions) with details on their computational efficiency. Further in this technical part, I talk about the programs developed for designing barcode regions emphasizing on their potentials and pitfalls. An introduction to approximate methods and metaheuristic techniques is also given in order to find the optimal solution for hard combinatorial problems. The last section of this chapter is about the analyses of DNA sequence data, here I talk about major sources of errors in DNA sequence data, sequencing chemistries and the available de-noising algorithms and programs. The main goal of this chapter is to give a background of the topic and introduce the reader with all the necessary material for understanding the rest of the chapters which present the main research work.

Second chapter of this thesis is about the importance of comparative study of several barcode markers and evaluation of quality of a given barcode region. In this chapter I present one of my publication where we published two formal quantitative indices designed for measuring the quality of a given barcode region, and, an *in silico* PCR application *ecoPCR* that can be used for comparing several barcode markers. I end this chapter by proposing some extensions to these quality indices taking into consideration the presence of errors in DNA sequence data.

Third chapter of my thesis is about designing new optimal barcode regions adapted to any particular application, especially for DNA metabarcoding for which many constraints should be taken into account. In this chapter, I present one of my publications on *ecoPrimers* program that I developed for designing application specific optimal barcode regions and which is capable of scanning large data sets like fully sequenced bacterial genomes. Further In this chapter I talk about selecting a set of minimum number of primer pairs from a given pool of primers such that most of the individuals from a given environmental sample are identified. In the same chapter I talk about the importance of using other target enrichment techniques than PCR and present a program that can design single primers to be used with DNA capture techniques.

The fourth chapter of this thesis is about DNA sequence data analysis in order to understand the potential sources of errors. In this chapter I present some preliminary results obtained from the analyses of sequences taken from snow leopards feces in order to determine its diet. The results of these analyses show that most of the errors are generated during PCR amplification step instead of sequencing process, contrary to what is generally believed. I conclude this chapter by suggesting that similar analyses should be performed on other data sets in order to see if same error behavior is observed. If it is so then algorithms should be developed for detecting and removing PCR generated artifactual sequences.

The fifth and last chapter is the discussion of results with some concluding remarks and perspectives of my work.

Acknowledgements

This dissertation and research work presented here is not one man’s effort. I am grateful to all of the people around me who were either directly involved in this work or gave me emotional support in the time of need.

First of all I am highly thankful to Eric Coissac for his kind supervision and help. Without his untiring efforts this thesis would have been impossible for me. Despite his busy schedule he was ever ready to help me cope with problems. I was always impressed by his knowledge of bioinformatics and his technical and programming skills and he has always been a source of motivation for me.

I am also thankful to Pierre Taberlet for incorporating me in his lab and for his valuable suggestions and comments. I am grateful to Christelle Melo de Lima for spending her valuable time reading my thesis and pointing out many corrections. I am also thankful to whole LECA team whose friendly attitude has made this place very comfortable to work.

I am highly thankful to the two reporters, Pierre Brezellec and Guy Perrier for reading this manuscript and evaluating my work.

I am thankful to my family for their moral and emotional encouragement during these years, especially my husband and my son who never let me down when I was stuck or lacked motivation. I am thankful to my friend Zunera Rais who read this manuscript in a short period of time and pointed out english mistakes.

Lastly I feel indebted to Higher Education Commission of Pakistan for providing all the financial support for this work and for making sure a continued support even when Pakistan was passing through financial crises.

To,

My beloved mother, whom I lost in this journey…

Contents

CHAPTER 1

Introduction

Our environment is a complex system where living organisms are functioning in interaction between them and with non-living physical factors around them. These organisms belong to a large variety of plants, animals, fungi and micro-organisms. The interesting feature about these organisms of all species inhabiting in this world is their uniqueness. This gives birth to what we call biodiversity. We see numerous types of plants and animals around us everyday. The plants found in plain areas are different from those found in mountains. Organisms vary in their structure, function and behavior and this variation depends on the part of environment in which they exist.

Along with the environment, the varieties of the organisms are a result of evolution through a very long period of time. During this period many species have become extinct, while numerous new ones have originated. Knowledge about organisms is indispensable for improving our understanding of the environment and for understanding the factors causing the extinctions and formations of new species. Humans have been in the continuous struggle of naming organisms and have been trying to organize life on earth into understandable categories. This categorization and naming process was the beginning of the identification and classification of organisms. An important concept related to categorization of individuals is species inventory. We define the concept of inventorying the species as the process of being able to distinguish groups of organisms present in an environment and assigning them to a possible taxonomic class. The concept of species inventory is somehow different from species identification, because identification also involves the discovery of new species. This thesis is mainly concerned with the techniques allowing the inventory of the species existing in our environment. In this work we have tried to show that species inventory is an important task and more efficient techniques are required to classify the largest part of species present on this earth.

1.1 Characterizing Species Diversity

The process of species inventory is of primary importance in many ecological studies especially for studying the diversity of life in an ecosystem. An ecosystem in which all plants, animals and micro-organisms are functioning together can be delimited by a geographical area of a variable size and the whole earth's surface can be described by a series of interconnected ecosystems. Within an ecosystem, all aspects of the environment interact and affect one another. Every individual affects the lives of those around him. An important concept related to ecosystem is the biodiversity which is defined as the degree of variation of life forms within itself. This variation of life forms can be described at all levels of biological systems including variety of habitats and processes occurring therein, variety at genetics level and at species level (and higher taxonomic levels) (Wilcox, 1984).

Based on the above definition, biodiversity can be defined on the following levels: ecosystem diversity, species diversity and genetic diversity. Ecosystem diversity refers to the diversity of a place at the level of ecosystems. Genetic diversity refers to the total number of genetic characteristics in the genetic makeup of a species and species diversity gives the number of species in an area and their relative abundance. Another very closely related term is "species richness" which is simply the number of different species in an area.

Species diversity can be measured by a common index called Simpson's Diversity Index (Simpson, 1949). This index measures the probability that two individuals randomly selected from a sample will belong to the same species. According to this index, if p_i is the fraction of all organisms which belong to the i^{th} species, then Simpson's diversity index can be formalized as:

$$D = \sum_{i=1}^{S} (p_i)^2$$

Campbell (2003) defined a fourth level of biodiversity as molecular diversity which is interpreted as the richness of molecules found in life. Examples include molecules forming the structures and metabolism of life, such as, amino acids forming proteins and sugars forming the backbone of nucleic acids and energy stores. According to him, life would not exist without any of these molecules. Although formally biodiversity is defined on these four levels, the most common interpretation is taken at the species diversity level and commonly biodiversity replaces the use of terms species diversity and species richness. Species is considered as the central unit of taxonomy and the characterization of the diversity of species living within an ecosystem is a major scientific interest in understanding the operations taking place thereof. Therefore, an unambiguous association of a scientific name to a biological entity is an essential step to build a reliable reference system of biological information. The concept of species inventory is quite old and different people in their time have been trying to group individuals in order to distinguish them and to classify them using different methods of biological classification and taxonomy. In the next section I will explain how classification has been done in the past and how it has evolved through the time.

1.2 Biological Classification And Taxonomy Through History

Biological classification is the process of grouping similar organisms together such that this arrangement shows the relationship among various organisms. Sometimes it is referred to as taxonomy, however taxonomy is a bigger concept and includes naming, identification and classification of organisms. The objective of classification in biology is to identify and make natural groups and to be able to describe organisms precisely by providing characteristics that can be useful in identification.

The classification system for different life forms was actually started by Aristotle. In his metaphysical work (Metaphysics Book VI), he published the first known classification of almost everything that existed in his time. In his book *scala naturae*, Aristotle gave the idea that the classification of a living thing should be done by its nature and not by superficial resemblance. This requires a close examination of specimen, observing their characteristics and noting which characters are constant and which are variable, as the variable characteristics may have been introduced due to environmental or accidental affects. Based on this idea, Aristotle studied animals and he classified them into two main groups; animals with blood (vertebrates) and animals without blood (invertebrates). He further divided animals with blood into live-bearing (mammals), and egg-bearing (birds and fish). Animals without blood were divided into insects, crustacea and testacea.

After Aristotle, major advances in classification were done by John Ray (1627-1705), an English naturalist who published important works on plants, animals and natural theology. He was the first person to introduce the term "animal species" and described more than 1800 plants and animals in his book *Historia Planturm* (Ray, 1686, 1688, 1704). However the modern classification system began after Carolus Linnaeus' publication *Systema Naturae* (1758). He grouped species according to shared physical characteristics and developed a hierarchical classification system for life forms in the 18th century which is the basis of the modern zoological and botanical classification and naming system for species. Linnaeus gave the concept that a group of organisms sharing a particular set of characteristics form an assemblage called taxon. Linnaean taxonomy is a rank based classification system which goes from general to specific. In the taxonomy of Linnaeus, there are three kingdoms namely animals, plants and minerals. These kingdoms are divided into phylums and classes, and they in turn into: orders, family, genera, and species. Species is the basic unit in Linnaean taxonomy and consists of individual organisms which are very similar in appearance, anatomy, physiology and genetics. However this definition of species is based on the modern concept of biological species introduced by Mayr (1942). In the time of Linnaeus, species concept was based only on morphology. The greatest innovation of Linnaeus is the general use of binomial nomenclature, which is the combination of a genus name and a second term, such that both names uniquely identify each species of organism.

The Linnaean system has proven robust and it is the only extant working classification system at present that has received universal scientific acceptance. However, over time, the understanding of the relationships between living things has changed (Ereshefsky, 2001). Linnaean scheme was based only on the structural similarities of the different organisms. However, morphological, physiological, metabolic, ecological, genetic, and molecular characteristics are all useful in taxonomy because they reflect the organization and activity of the genome. Therefore, after the publication of Charles Darwin *On the Origin of Species* in 1859, it was accepted that classification should reflect Darwin's theory of common descent (Hey, 2005). In this book Darwin (1859) presented convincing evidence that life had evolved through the process of natural selection. This theory states that all species of life have descended over time from a common ancestor. The immediate impact of Darwinian evolution on classification was negligible, however with time taxonomists started accepting the concept of evolution. As a consequence classification since Linnaeus has incorporated newly discovered information and more closely approaches a natural system by explaining that the similarity in forms and characteristics is actually an evolutionary descent relationship. People started accepting that the similarity between organisms is not a coincidence; organisms actually inherited these traits from the same common ancestor. In general, the greater the resemblance between two individuals, the more recently they diverged from a common ancestor. With the acceptance of Darwin's theory of evolution, scientists started to represent classification in the form of tree of life and a concept of reclassifying the life appeared. The first fossil record found at that time belonged to dinosaurs and based on Darwin's theory of evolution, birds were tied to this fossil record saying that birds are descendants of dinosaurs. However, not many fossil records were found and due to very limited knowledge of the fossils at the time, scientists were not very successful at drawing specific inferences about the ancestors of modern groups and Darwin could only show the relationship between living organisms.

1.3 Common Methods For Classifying Organisms

Organism classification methods can broadly be divided in two groups: methods based on physical traits called phenetics classification and the methods based on evolutionary relationships like phylogenetics classification and evolutionary systematics. Both types of methods aimed at designing objective and biologically meaningful ways of classifying organisms or, alternatively, to invent a procedure for approximating or estimating the true natural relationships among organisms. Brief description of both of these classification methods is given next.

1.3.1 Classification Based On Physical Traits

The most well know method of classification based on physical traits is phenetics classification. This system is closely related to numerical taxonomy. Numerical taxonomy is a biological classification system that deals with the grouping of organisms by numerical method based on their characteristics. The concept was developed by Sneath and Robert (1973). The branch of numerical taxonomy was divided into two fields called phenetics (classifications based on patterns of overall similarity) and cladistics (based on the branching patterns of evolutionary history of the taxa). However, in recent years, numerical taxonomy and phenetics are used synonymously despite their original distinction.

Phenetic systems of classification started with Carolus Linnaeus himself. All the taxonomy in the beginning was based on this method. This method relies on similar and dissimilar features present in organisms or other observable traits without including phylogeny, evolutionary and other related aspects. This method emphasizes on numerical analyses of an observed set of phenotypic characteristics and includes various forms of clustering and ordination. These clustering and ordinations are important to reduce the variation displayed by organisms to a manageable level. Although it seems to be a straightforward task to measure a large number of traits of organisms and then assess the degree of similarity among them, in practice it is not so simple. This is because one needs to make decisions whether, some traits are more important than others, and, whether a group of traits that are all direct responses to a single selective pressure should be given the same weight as traits influenced by different selective pressures.

The technique of phenetics has largely been superseded by cladistics approach for research into evolutionary relationships among species. However, one important phenetic method called neighbor-joining (Saitou and Nei, 1987), which is a bottom-up clustering method for the creation of phenetic trees, is still in use as a reasonable approximation of phylogeny when more advanced methods like Bayesian inference are computationally expensive.

1.3.2 Classification Based On Evolution

Phenetic analysis do not distinguish between traits that are inherited from an ancestor and traits that evolved as new in one or several lineages. Consequently results based on phenetic analysis can be misleading. Modern classification is based on classifying organisms based on evolutionary descent, rather than physical similarities. This type of classification makes use of molecular techniques to find out the variations in genotypes. One of the most important and well-known method based on molecular techniques is phylogenetics systematics (or cladistics). This type of classification is concerned with grouping individual species into evolutionary categories by making use of the structure of molecules to gain information on an organism's evolutionary relationships. Cladistics approach classifies individuals into groups called clades which consist of an ancestor organism and all its descendants. Cladistics originated in the work of the German entomologist Willi Hennig who referred to it as "phylogenetic systematics" (Willi, 1979). This method is realized with the depiction of cladograms which show ancestral relations between species. Cladistic analysis has a strict rule that new species arise by bifurcations of the original lineage and hence the lineage always splits in two.

Another classification method based on evolutionary insight is called evolutionary systematics or Darwinian classification. This method seeks to classify organisms using a combination of phylogenetic relationship and overall similarity (Mayr and Bock, 2002). Evolutionary systematics differs from cladism in that cladism only maps phylogeny where each taxon must consist of a single ancestral node and all its descendants and hence only two branches are possible (Grant, 2003). A simple example of the difference between two approaches could be that birds and crocodilians diverged from the same ancestral reptilian line. A cladist would insist that these "sister groups" be placed in the same taxon, even though the amount of change from the common ancestor is much greater for birds than it is for crocodiles. An evolutionary taxonomist would suggest that the large number of similarities between crocodilians and reptiles would justify grouping them within the same general taxon, while placing birds in a separate taxon due to the large number of unique characteristics possessed by members of this group. Apart from this difference of inclusion of similarity into classification, the rest is same; both techniques use the information from evolutionary history to classify individuals.

Molecular data used to gain insight into an organisms evolutionary history include protein and DNA sequences. Closely related organisms generally have a high degree of similarity in the molecular structure of these substances, while the molecules of organisms distantly related usually show a pattern of dissimilarity. Molecular phylogeny uses such data to build a "relationship tree" that shows the probable evolution of various organisms.

The most common approach used for analysis of genomes of various organisms is the comparison of homologous sequences for genes using sequence alignment techniques to identify similarity. One of the recent applications of molecular phylogeny is DNA barcoding, where the species of an individual organism is identified using small sections of DNA. DNA barcoding is a smart and efficient method of choice to discriminate several individuals and to assign them to a possible taxonomic class. This technique will be discussed further in detail in the next section.

1.4 DNA Barcoding

As we advance in understanding cellular DNA and the building blocks of species, we may be able to define organisms more precisely by making use of the emerging fields of DNA sequencing and DNA barcoding. The advent of DNA sequencing has significantly accelerated biological research and discovery. DNA sequencing is the process of determining the order of the nucleotide bases; adenine, guanine, cytosine, and thymine in a molecule of DNA. Knowledge of DNA sequences has become indispensable for basic biological research. This technique has made it possible to use DNA sequences as a major source of gaining new information for advancing our understanding of evolutionary and genetic relationships (Hajibabaei *et al.*, 2007). DNA sequencing can also be used for species identification through the technique of DNA barcoding. DNA barcoding is a molecular method to identify species through the analysis of variability of a single standard DNA region. It is an old concept. Carl Woese was the first person to use nucleotide variations in rRNA to discover Archea in 1977 (Woese. *et al.*, 1990). He recognized that sequence differences in a conserved gene, ribosomal RNA could be used to infer phylogenetic relationship. However the term DNA barcodes was first used by Arnot *et al.* (1993) in their article on using Digital codes from hyper-variable tandemly repeated DNA sequences. But this publication did not receive much attention from scientific community. The actual golden period of DNA barcoding started in early 2000, after the publication of Floyed on using molecular barcodes for soil nematodes identification (Floyd *et al.*, 2002) and Paul Herbet's publication on biological identification through DNA barcodes (Hebert *et al.*, 2003a). The term DNA barcodes is used as an analogy with the Universal Product Codes on manufactured goods.

It is important to say that despite of some popular misconceptions, the goal of DNA barcoding is neither to determine the tree of life nor to carry out phylogenetic studies. The goal of DNA barcoding is also not molecular taxonomy, as it is not intended to replace classical taxonomy. Its purpose is to carry out species identifications without involving the expert's knowledge and doing so in a rapid and inexpensive manner.

1.4.1 The Barcoding Principle

The principal concept behind DNA barcoding is quite simple *i.e.* a short DNA sequence can distinguish individuals of different species. More explicitly it can be stated that through the analysis of the variability in a single or in a few molecular markers, it is possible to discriminate biological entities. The barcoding method is based on the assumption that genetic variation between species exceeds than that which is within the species. This is because some DNA regions evolve more rapidly than others between species, and, vary to a minor degree among individuals of the same species, giving rise to a higher genetic variation between species and relatively less variation within species (Hebert *et al.*, 2003a). This is true for mitochondrial DNA. Most eukaryote cells contain mitochondria, and mitochondrial DNA (mtDNA) has a relatively fast mutation rate, which results in significant variation in mtDNA sequences between species and, in principle, a comparatively small variation within species. This is the reason that, DNA sequences of a suitable length can provide an unambiguous digital identifying feature for species identification.

In this context Paul Hebert proposed a 648 bp region of the mitochondrial cytochrome c oxidase subunit I (*COI*) gene as a potential barcode for animal identification (Hebert *et al.*, 2003b). A number of studies have shown that for higher animals, the variability at the 5′ end of *COI* gene is very low (about 1 to 2 percent) but even between closely related species it differs by several percents, making this an ideal region to be set as the standard for animals DNA barcoding. In some groups, *COI* is not an effective barcode region like in plants because of much slower evolution rate of *COI* gene in higher plants than in animals (Kress *et al.*, 2005) and hence a different standard region should be sought and agreed on. However, the idea is that in all cases, DNA barcoding uses a short and standard region that enables cost-effective species identification.

At the beginning, the original idea was to apply DNA barcoding to all metazoa by using mitochondrial marker *COI*. Rapidly, the idea was extended to flowering plants (Kress *et al.*, 2005), and fungi (Min and Hickey, 2009) and now DNA barcoding initiative can be considered as a tool suitable for the whole tree of life. The development of DNA barcoding as standard for species identification is being done by Consortium for the Barcode of Life (CBoL).

1.4.2 Role Of The Consortium For The Barcode Of Life

Today DNA barcoding is a well-established research field and there is a consortium called CBoL that has taken charge of the development of DNA barcoding as a global standard for the identification of biological species. CBoL has provided a protocol for DNA barcoding. According to this protocol barcoding begins with the collection of specimen followed by obtaining DNA barcode sequence from a standard part of the genome of these specimen which requires laboratory analysis (*e.g.*DNA extraction, Polymerase Chain Reaction aka PCR and the sequencing of amplified region). This barcode sequence obtained from unknown specimen is then compared with a library of reference barcode sequences derived from individuals of known identity. The specimen is identified if its sequence has sufficient similarity with one in the library, otherwise a new record is added which can be considered as the new barcode sequence for a given species (new haplotype or a geographical variant). One of the most important component of DNA barcoding is to maintain a public reference library of species identifiers which could be used to assign unknown specimens to known species. There are three general purpose public databases which contain published DNA sequences. These are GenBank,[1] EMBL[2] and DDBJ.[3] However, the quality of the sequence data in these databases is not always perfect. This could be because of sequencing errors, contaminations, sample misidentifications or taxonomic problems (Harris, 2003). In this context CBoL has taken an initiative to build a new database especially dedicated to DNA barcoding. This database system called BOLD (Barcode of Life Data Systems)[4] is designed to record DNA sequences from several individuals per species along with complete taxonomic information, place and date of collection, and specimen images (Ratnasingham and Hebert, 2007).

1.4.3 The Choice For A Suitable Barcode Loci

Theoretically speaking a barcode marker consists of two conserved regions flanking a central variable region (Ficetola *et al.*, 2010). The conserved regions actually work as primers for PCR amplification and the central variable part allows species discrimination. To diagnose and define species by their DNA sequences on a large and formalized scale, we need to identify genome regions that fulfill certain properties; the chosen locus should be standardized (in order to develop large databases of sequences for that locus), it should be present in most of the taxa of interest, should be short enough to be easily sequenced

[1] http://www.ncbi.nlm.nih.gov

[2] http://www.ebi.ac.uk/embl

[3] http://www.ddbj.nig.ac.jp

[4] http://www.barcodinglife.org

with current sequencing technologies (Kress and Erickson, 2008) and provide a large variation between species yet a relatively small amount of variation within a species (Lahaye *et al.*, 2008). Based on these properties several loci have been suggested, however the most well-known is *COI* gene (Hebert *et al.*, 2003b). In addition to *COI*, several regions of *RNA* genes like *12S* (Kocher *et al.*, 1989) or *16S* (Palumbi, 1996) *rDNA*, and non-coding chloroplastic regions such as the *trnL* intron (Taberlet *et al.*, 2007), some intergenic regions as *trnH-ps* (Kress *et al.*, 2005) and the gene of the ribulose-bisphosphate carboxylase (Hollingswortha *et al.*, 2009) are also being used in different groups of animals, plants and fungi.

1.4.4 DNA Barcoding Types

DNA barcoding can be divided into two main types according to its applications in different fields. These two types are called DNA barcoding *sensu stricto* and DNA barcoding *sensu lato* (Valentini *et al.*, 2009). The *sensu stricto* barcoding, *i.e.* barcoding in the strict sense is the standard barcoding approach as defined by CBoL which emphasizes on the identification of the species level using a single standardized DNA fragment. This approach is more adapted by taxonomists. However, taxonomists are not the only potential users of DNA barcoding. DNA barcoding has diverse applications and can be useful for scientists from other fields including ecology, biotechnology, food industries, animals diet and forensics. DNA barcoding can be of great help in conservation biology for biodiversity surveys, for reconstituting past ecosystems by studying fossil soils and permafrost samples, for studying the molecular signature of bacteria from soil ecosystems which is an important tool to study microbial ecology and bio-geography (Zinger *et al.*, 2007) and it could also be applied for the analysis of stomach contents or fecal samples to determine food webs.

However, all these applications come in the category of DNA barcoding *sensu lato* which corresponds to DNA based taxon identification using diverse techniques that lie outside the CBoL approach. The difference between the two approaches derives mainly from different priorities given to the criteria used for designing the molecular markers. We refer to *sensu lato* barcoding approach as DNA *metabarcoding* (Pompanon *et al.*, 2011) or *environmental barcoding* which could be defined as simultaneous identification of multiple species from environmental samples using high throughput sequencing techniques. This approach will be discussed in more detail in the next section.

1.4.5 Applications Of DNA Barcoding

Our planet is populated by millions of species and their identification is an important, but at the same time, a not so easy task. Historically, species have been described and characterized on the basis of morphological criteria. Since Carl Linnaeus's classification system, about 1.7 million species have been formally described by taxonomists which according to an estimate comprises of 20% of eucaryote life on earth, but it is largely accepted that this number probably represents only a small fraction of the real biodiversity present on the earth (Vernooy *et al.*, 2010).

There are, however, limitations to relying solely or largely on morphology in identifying and classifying life's diversity. Morphological characterization based on visible traits is the most natural and intuitive method that distinguishes species but at the same time it is a complex process and most taxonomists can only specialize in a single group of closely related organisms. As a result, a multitude of taxonomic experts may be needed to identify specimens from a single biodiversity survey. Moreover finding appropriate experts and distributing specimens can be a time consuming and expensive process. Thousands of expert taxonomists are required to identify life on earth even if we consider that morphological identification method is reasonably reliable, but the reality is different. Moreover morphological identification method can be misleading in some individuals if somehow the particular trait of interest is changed in response to environmental factors (Hebert *et al.*, 2003a) such as in the case of cryptic species.

Although DNA barcoding is potentially used for specimen identification, it is especially useful in the cases where traditional morphological methods fail, for example identification of eggs and immature forms (Zhang *et al.*, 2004) including many other examples. In the next section, I discuss some of the well known examples where barcoding either improved the results obtained from morphological analysis, or where morphological analysis was difficult to use.

Identification Of Cryptic Species

Cryptic species is a group which satisfies the definition of species by being reproductively isolated from each other but their morphology is very similar and in some cases they are identical (Knowlton, 1993). Mostly insect parasitoids contain cryptic species. Insect parasitoids are a major component of global biodiversity and they are known to be a major cause of mortality for many host insect species. Thus they strongly affect the population dynamics of their hosts. Tachinid Fly (*Belvosia nigrifrons*) is known to be a cryptic species. The larvae of most Tachinidae fly species are parasitoids of insect larvae of butterflies and moths.

Morphological based identification of such cryptic groups is quite difficult because of the very large number of morphologically similar species however DNA barcoding has proven to be quite successful in their identification. Smith *et al.* (2006) worked on almost 3,000 *Belvosia* flies which were reared from caterpillars. The 3,000 flies were grouped into 20 species by an expert fly taxonomist. Out of these 20 species, 17 were thought to be host specific, while, 3 were considered to be more generalist in host selection. *COI* gene was used as barcode to determine whether the 20 species could be identified by their DNA barcodes. The barcoding results clearly discriminated among 20 species further proving that the 3 assumed generalist species were arrays of three, four, and eight cryptic species, each using a set of specific hosts like other 17, and thus raised the total number of species from 20 to 32.

Recently Janzen *et al.* (2009) used this method for the inventory of caterpillars. Author says that barcoding has been found to be extremely accurate during the identification of about 100,000 specimens of about 3500 morphologically defined species of adult moths, butterflies, tachinid flies, and parasitoid wasps, and, less than 1% of the species had such similar barcodes that a molecularly based taxonomic identification was impossible. Moreover no specimen with a full barcode was misidentified when its barcode was compared with the reference library.

Forensic Science

DNA barcoding has proven to be a powerful tool to help in the identification of species for forensic purposes. It has been successfully used for monitoring illegal trade in animal by-products where identification through morphological characteristics might not always be possible. For examples hair of Eurasian badger (*Meles meles*) have been found by Roura *et al.* (2006) in shaving brushes made in different European countries where this species is considered a protected species. The population of the Tibetan Antelope (*Pantholops hodgsonii*) has recently declined dramatically due to the illegal trade in its wool. Lee *et al.* (2006) have successfully shown through DNA testing that some shawl samples of sheep wool (*Ovis aries*), cashmere from the Kashmir goat (*Capra hircus*), and pashmina from the Himalayan goat (*Capra hircus*) actually contained wool from *Pantholops hodgsonii* as well.

DNA profiling is a technique employed by forensic scientists to assist in the identification of individuals by their respective DNA profiles. It is used in, for example, parental testing and criminal investigation. Currently the DNA Shoah Project is under process which is a genetic database of people who lost family during the Holocaust. The database is aimed to serve to reunite families separated during wartime.

Biodiversity Assessment

Understanding how populations and communities are structured and what triggers global biodiversity patterns, could be an essential point to predict future response of species diversity to environmental changes and to define efficient conservation strategies (Niemelä, 2000). Current knowledge about global biodiversity is quite partial, and mostly restrained to some well-studied areas and taxa. Global and extensive biodiversity assessment is restricted by the difficulty to aggregate data from different studies, due to absence of standardized methodology and approaches (Whittaker, 2010). A tool allowing a standardized biodiversity assessment (including all taxa and areas) is thus required to be able to have a global vision on biodiversity. In this context DNA barcoding can be helpful in the assessment of biodiversity. Although morphological identification is possible for assessing biodiversity of well-known ecosystems but in ecosystem of tropical regions with high species richness, the use of morphological method is unrealistic to identify all individuals within a given time period. Moreover the morphological method is difficult to apply in ecosystems where access is not easy, for example, to study the microbial biodiversity in deep sea (Sogin *et al.*, 2006). DNA barcoding could allow biodiversity assessment through the identification of taxa from the traces of DNA present in environmental samples such as soil or water. Moreover, with barcoding, large scale studies become easily possible because DNA barcoding allows simultaneous identification of a large number of species (Valentini *et al.*, 2009) from a given environmental sample and hence speeding up the assessment process. DNA barcoding can also help in measuring the diversity of meio- and micro-fauna and flora (Blaxter *et al.*, 2005) which are a key to the functioning of ecosystems because macro-organisms rely on them for their existence. However, because of small size of these meio- and micro-fauna and flora, facile visual identification even through a light microscopy may not always be possible. But DNA barcoding may permit rational access to these organisms by making use of water or soil samples.

DNA barcoding can also complement the biodiversity indices such as species richness and Simpson's index by integrating the definition of MOTU (Blaxter *et al.*, 2005, Floyd *et al.*, 2002). This could be achieved by estimating these indices based on molecular operational taxonomic units (MOTU) detected using the barcoding approach where the relative abundance of each type of DNA sequence (MOTU) replaces the classical relative abundance of each species estimated from the number of individuals, however such an approach can create a bias for larger number of species (Blackwood *et al.*, 2007). Such a bias could be generated by overestimating the biodiversity if there occur many MOTU for a single species or by underestimating the diversity if a single MOTU is found for many different species (Hickerson *et al.*, 2006). This approach is often used for estimating microbial diversity (*e.g.* Herrera *et al.*, 2007, Vicente *et al.*, 2007).

Paleoecology

Knowledge about past species and their environmental and climatic variation can play an important role in projections of future climate change effects on species and ecosystems (Boessenkool *et al.*, 2010). However, such analysis depend on species identification from remains of past animal and plant communities that exist in the form of fossils. Species level identification from these low preserved fossil records through morphological identification may be very difficult or almost impossible. But recent advances in species identification techniques making use of high throughput sequencing and DNA barcoding are proving quite helpful in gaining knowledge about past species, and in reconstructing the past ecosystem. A study done by Willerslev *et al.* (2007) on samples collected from 450,000 year old silty ice extracted from the bottom of the Greenland ice cap revealed that southern Greenland was covered by a forest at that time, composed of trees of the genera *Picea*, *Pinus* and *Alnus* as in the forests found in southern Scandinavia today. Another study based on the DNA analysis of 11,700 years old rodent middens from the Atacama Desert in Chile was done by (Kuch *et al.*, 2002). In this study DNA was extracted from old rodent middens, and, chloroplast and animal mitochondrial DNA (mtDNA) gene sequences were analyzed to investigate the floral environment surrounding the midden, and the identity of the midden agents. The study revealed that this past environment was more productive with 13 plant families and 3 orders that no longer exist today. The environment was more diverse and much more humid, with a fivefold decrease in precipitation since that time. These and other similar studies (Hofreiter *et al.*, 2000, Willerslev *et al.*, 2003) reveal that the association of ancient DNA with the barcoding concept offers new and promising opportunities to reconstruct past environments.

Diet Analysis

Molecular identification of animal or plant species in fresh and degraded products (*e.g.* food, feces, hair and other organic remains) has become a very important issue in both conservation biology and food science. The study of feeding ecology is vital both for constructing food webs and taking measures to conserve endangered species. A food web is defined as a graphical description of feeding relationships among species in an ecosystem. Study of food webs to determine the diet and feeding behavior of species present in a given environment can improve our understanding of the functioning of the ecosystem. Reliable assessment of food web structure depends on correct understanding of each individual trophic interaction. To date, the construction of food webs has largely been based on morphological characteristics of species. However, Kaartinen *et al.* (2010) working on gall wasps and leaf-miners have shown that the food web based on morphological characters contained 25 host species, 51 species of parasitoids and 5 species of *Synergus inquilines*, whereas the web improved with molecular characters included 58 parasitoid species and 6 *Synergus* MOTUs.

Also precise knowledge of the diet of endangered species can be helpful in identifying key environmental resources for designing reliable conservation strategies. DNA barcoding makes it possible to establish the diet of an individual from its feces or stomach contents. This is important because the prey choice by predators in the field cannot always be established using direct observation. DNA barcoding is particularly useful in diet determination when the food is not identifiable by morphological criteria, such as in the case of spider which usually ingests only liquid and soft body tissues from prey species (Agustí *et al.*, 2003), or when the diet cannot be deduced by observing the feeding behavior *e.g.* deep sea invertebrates and diatom-feeding krill (Passmore *et al.*, 2006). Moreover for certain predator species, prey identification involves sampling procedures that are disruptive for the predator, such as stomach flushing (Jarman *et al.*, 2002). In such cases DNA barcoding is a more descent method especially for those species which are already endangered like snow leopard.

1.4.6 Multiple Species Identification And Limitations Of Standard Barcoding

The applications of DNA barcoding described above can be divided into two categories; those who make use of standard barcoding and are based on single species identification and those who make use of less restrictive approach of barcoding i.e. DNA metabarcoding and identify multiple taxa from a single sample. For example, identification of cryptic species belong to single species identification, but reconstructing past environment belongs to metabarcoding where the analysis of a single water or soil sample can give information about different kinds of plants, animals and microbial species. Although single species identification is the historical fundament of DNA barcoding, and, there are many situations where DNA based single species identification can help the taxonomists to solve important ecological questions, yet DNA barcoding has much more potential than this. DNA barcoding can be successfully used for simultaneous multiple species identification from a single environmental sample. Environmentalists are usually more interested in this approach, and they have a broader view, corresponding to the use of any technique of DNA analysis for identification of taxa (Valentini *et al.*, 2009) with the objective to identify a large set of taxa present in an environmental sample even if the identification at species level is not possible. However, due to inherent problems of *sensu stricto* barcoding, this approach is of limited use for metabarcoding application.

Since the standard DNA barcoding as defined by CBol uses a single and pure specimen, DNA extracted from this specimen is of high quality and enough of DNA is available for analysis. Thus use of standard barcode markers like *COI* for animals and *rbcL* or *matK* for plants is appropriate. On the other hand, the DNA present in environmental samples is mostly degraded and the amount of DNA extracted is also very less. The main limitation in using standard markers lies in the length of the sequences used which are usually > 500 bp (Hebert *et al.*, 2003b). This long length prevents the amplification of degraded DNA. Unfortunately, many potential DNA barcoding applications in the field of ecology can only be based on degraded DNA. This is the case for all environmental samples where the target is DNA from dead animals or dead parts of plants or DNA taken from feces or from permafrost samples. In all these cases, it is difficult to amplify DNA fragments longer than 150 bp from such samples (Deagle *et al.*, 2006). Another limitation of *sensu stricto* barcoding is that this approach insists on species level identification, but with environmental samples species level identification may not be possible because of the low resolution of short markers. In such cases identification at any taxon level is acceptable given that most of the taxa present are well discriminated and identified.

1.5 DNA Metabarcoding

In the above sections we have shortly talked about DNA metabarcoding and environmental samples. In this section we will clearly define these terms and talk about them in detail. First we will see what is an environmental sample because metabarcoding is based on the use of such samples.

An environmental sample is a mixture of some organic and inorganic materials taken from environment, for example a water sample taken from deep sea to study biotic communities or soil sample taken from an ecosystem to study species diversity or feces sample to study diet of certain animal species. These type of samples can contain live micro-organisms or small macro-organisms such as nematodes or springtails and remains of dead macro-organisms present around the sampling site. This DNA can be extracted and albeit partially degraded, short sequences can be amplified and sequenced. Soil and deep sea water samples represent a potential information source about all organisms living in them, and these samples can be used to have an overview of organisms' diversity by using metabarcoding approaches.

At the beginning most of the studies done on environmental samples were focused on microbial communities (Herrera *et al.*, 2007, Vicente *et al.*, 2007, Zinger *et al.*, 2008). In this case DNA sequences of several hundreds of base pairs can be retrieved because DNA of good quality is extracted from live microorganisms. However, environmental samples can also be used for characterizing the diversity of macro-organic species such as plants or animals in an ecosystem, where DNA comes from dead macro-organisms, and in most cases it is highly degraded. In this case only short sequences can be amplified.

DNA metabarcoding or environmental barcoding corresponds to the identification of any taxonomic level (not restricted to species level) using any suitable DNA marker (and not just the standardized markers). Thus the identification of genera or families, from an environmental sample using a suitable short DNA fragment that has not been recognized as the standardized barcode, can be considered as DNA metabarcoding. Metabarcoding requires DNA extraction from an environmental pooled sample, PCR amplification from a mixture of degraded DNA samples, sequencing large numbers of DNA barcodes using high-throughput sequencing techniques and the analysis of this huge amount of sequence data. DNA metabarcoding, thus has the potential to provide the accurate measures of genetic richness in the quantitative samples taken at each sampling point.

1.5.1 DNA Metabarcoding With New Sequencing Techniques

Classical barcoding system is based on Sanger sequencing approach (Sanger *et al.*, 1977) and can target single specimens. Sanger sequencing yields a read length of $800 - 1000$ bp. This approach is not feasible for environmental samples where mixtures of organisms are under investigation. However, recently next-generation sequencing systems have become available (Hudson, 2008, Schuster, 2008). These new sequencing technologies can aid in directly analyzing biodiversity in bulk environmental samples through their massively parallelized capability to read thousands of sequences from mixtures (Hajibabaei *et al.*, 2009).

This new, fast and cheap DNA sequencing in short segments is the most innovative recent development. Several new sequencing techniques have been developed which are based on methods that parallelize the sequencing process allowing the simultaneous sequencing of thousands or millions of sequences at once (Church, 2006, Hall, 2007). These sequencing methods include the 454 implementation of pyrosequencing, Solexa/Illumina reversible terminator technologies, polony sequencing and AB SOLiD. The typical read length of 454

GS FLX/Roche is 500 bp, for Solexa/Illumina it is 100 bp, and for polony sequencing and AB SOLiD it is $25 - 50$ bp. The enormous amount of relatively long sequences produced by 454 GS FLX/Roche and Solexa/Illumina, make these new sequencers suitable for environmental barcoding studies where scientists have to deal with complex samples composed of a mixture of many species *e.g.* deep sea biodiversity (Sogin *et al.*, 2006) and diet analysis (Shehzad *et al.*)(submitted).

1.5.2 Barcode Designing For Metabarcoding Applications

Having talked about the usability of DNA metabarcoding and its vitality in ecological studies, now the question arises how can we successfully use this approach? Considering the broader view of metabarcoding and its applications in the field of biodiversity, forensics, diet analysis and paleoecological studies which are based on the analysis of environmental samples, it is easy to conclude that standard barcode markers as defined by CBoL are not suitable for metabarcoding studies. In order to perform DNA metabarcoding effectively the first step of a metabarcoding study should be the selection of best DNA regions to be used as barcodes considering the aim of the study. It has been suggested that shorter barcoding markers should be used (Taberlet *et al.*, 2007). However before talking about the design of barcode markers, we need to know what are the properties of an ideal barcode marker.

According to both theoretical and experimental points of view, an ideal barcode marker should fulfill the following properties (Valentini *et al.*, 2009).

- The DNA region selected as barcode should be nearly identical among individuals of the same species, but different between species, giving it a strong discriminating power.
- It should be standardized as defined by CBoL so that the same DNA region could be used for different taxonomic groups.
- The target DNA region should contain enough phylogenetic information *i.e.* the level of divergence between these reference sequences reflects the level of divergence between actual species so that unknown or not yet “barcoded“ species could be easily assigned to their respective taxonomic group (genus, family, *etc.*).
- It should be flanked by two highly conserved regions from one species to another to allow amplification of the fragment by PCR in as many species as possible, thus ensuring a good taxonomic coverage. This is particularly important when using environmental samples, where each extract contains a mixture of many species

to be identified at the same time. This property is also important for simplifying PCR amplification conditions to reduce disequilibrium in amplification amongst the different DNA templates and to avoid the production of possible chimeric products.

- The target DNA region should be short enough to allow amplification of degraded DNA. Usually, DNA regions longer than 150 bp are difficult to amplify from degraded DNA.

Taking into account the scientific and technical contexts, the various categories of users (*e.g.*taxonomists, ecologists, *etc...*) will not give the same priority to the five criteria listed above. The first three criteria are the most important for taxonomists (DNA barcoding *sensu stricto*), whereas ecologists working with environmental samples will favor the last two criteria. Unfortunately there exist no such markers with these properties suitable for metabarcoding applications. Moreover different metabarcoding applications may need different barcode markers. In the following subsection we will see that how can we efficiently design barcode markers specific to a particular application considering the aims of the study.

Barcode Design Workflow

In order to design the barcodes which are most relevant to a particular study, we can make use of the large public databases of sequences that exist today (Ficetola *et al.*, 2010). We can perform a database search to extract sequences that belong to a targeted organism or taxa. Mostly sequences are downloadable from *GenBank*, *EMBL* or *DDBJ*. In order to search the relevant sequences for a particular study, for example from *NCBI's GenBank*, *BLAST* program (Altschul *et al.*, 1997) can be used. For downloading the sequences, *NCBI* has provided the utility of *Entrez* (Wheeler *et al.*, 2006) which is web-based search and retrieval system for major databases. Once we have our target sequences as input, we can identify conserved regions shared by these sequences in order to design barcode markers. Finally the selected conserved regions need to be checked against certain criteria to be used as PCR primers and eventually as barcode markers.

In all these steps, finding conserved regions (also called repeated patterns) is the most challenging task. It is an important and widely studied problem in computational molecular biology and there exist a number of different computer science techniques to find such regions.

1.6 Finding Conserved Regions

Finding biologically meaningful segments *e.g.* conserved segments is an important line of research in sequence analysis. In biology, conserved sequences are defined as those DNA regions which are highly similar or identical throughout a large number of taxa. This conservation may be a consequence of functional, structural or evolutionary relationships between sequences. In the case of cross species conservation, this indicates that a particular sequence may have been maintained by evolution despite speciation. Conserved regions in DNA or protein sequences are strong candidates for functional elements, and so the development and comparison of appropriate methods for finding these regions is very important.

The conserved regions can be divided into two types, one which are strictly identical called strict motifs and the others which are similar but not strictly identical, called approximate motifs. The problem of finding these two types of conserved regions among a set of DNA, protein or amino acid sequences is called the problem of motif finding. Besides designing barcode markers, motif finding applications arise when identifying shared regulatory signals within DNA sequences or shared functional and structural elements within protein sequences. Due to the diversity of contexts in which motif finding is applied, several variations of the problem are commonly studied (Hu *et al.*, 2005). Motif finding or locating conserved regions is possible through informatics techniques which can be mainly divided into two categories. The first is the empirical method of locating conserved regions from a sequence alignment and the second is without alignment using combinatorial or probabilistic techniques. These techniques will be discussed in detail in the next section, however before talking about the methods available for finding conserved regions, it is important to know about some computer science considerations.

1.6.1 Some Computer Science Considerations

Any method used to accomplish a certain task (*e.g.* inferring barcode region from a set of sequences) corresponds to an algorithm. An algorithm is a set of well-defined rules or procedures that is designed to systematically solve a certain kind of problem in a finite number of steps. A certain number of properties are associated with algorithms. From these properties, the one showing the relationship between the size of the input data and the computational capacity needed by the algorithm to find its solution, is the most important. This property is named as complexity. It determines our actual capacity to compute the solution. We define two types of complexities:

- the complexity in time, that links the input size and the computational time.
- the complexity in space that links the input size and the amount of memory needed to achieve the calculus.

Complexity is expressed as a function noted by O, also called big-O notation. For example, a complexity in $O(n)$ indicates that the computational capacity grows linearly with the size of data, a complexity in $O(n^2)$ indicates that the computational capacity grows quadratically with data size. Complexity is calculated in the "worst case". But sometimes with certain data sets, the effective computational time can be faster than the one predicted by complexity function. Also for some algorithms "mean case" complexity can also be estimated on real data. If the complexity of an algorithm is too high, we can define a heuristic. A heuristic is a computational method which begins with only an approximate method of solving a problem within the context of some goal, for computationally difficult problems.

As previously explained, looking for conserved regions is the same problem as looking for repeats in sequences. Since we are looking for conserved regions among a set of sequences to be able to design primers, thus the properties of such regions are constrained by PCR experiment. We know that some differences are tolerated between the primer sequence and the matrix sequences. From this assumption, we can define the kind of repeats we are looking for. We are working on DNA sequences that can be assimilated to a string, where:

Definition 1. A string τ is an ordered set of symbol s_i where i is the position of the symbol in τ.

$$i \in [0, length(\tau)[$$

Each symbol s belongs to a particular alphabet σ

$$\forall\, i \in [0, length(\tau)[\;\Rightarrow\; s_i \in \sigma$$

For DNA sequences; $\sigma = \{a, c, g, t\}$. All contiguous subset of positions on a string are defined as a word.

On a string τ, strict repeats can be defined as following:

Definition 2. Two words of length l on two positions m and n of string τ are strictly identical (or strict repeats) if they satisfy following condition:

$$\tau[m+i] = \tau[n+i] \mid 0 <= i < l$$

Here $\tau[j]$ means the letter on position j of string τ. We will represent this strict repeat as $\rho_{m,n}$. In a similar way we can define repeats occurring in more than two positions.

In the preceding definition, no errors (differences) are allowed between copies of the repeats. But we have explained that PCR experiment tolerates some errors in conserved regions. We want to tolerate errors, such that, on conserved region where we locate the primer, no more than e errors are present between all the copies of the conserved regions. This divergence between two words can be measured by a *Hamming distance*.

Definition 3. The Hamming distance d_H between two words ρ_1 and ρ_2 of length l is the count of positions i of ρ_1 and ρ_2 where

$$\rho_{1,i} \neq \rho_{2,i}$$

Thus in order to design PCR primers we have to find non-strict (approximate) repeats where all words ρ included in this repeat, of length l equal to the size of the searched primers, have a hamming distance $d_H \leq e$ between each of its copies. There exist a number of methods to find strict repeats and approximate repeats each suffering from some limitations.

1.6.2 Locating Conserved Region With Multiple Sequence Alignment

Sequence alignment is a way of arranging the sequences of DNA, RNA or protein to identify regions of similarity. Sequence alignment can be performed in pairs, where two query sequences are aligned at a time, a scheme called pairwise sequence alignment. Alignment can also be performed for more than two sequences through multiple sequence alignment which is an extension of pairwise alignment to incorporate more than two sequences at a time. Multiple alignment methods try to align all of the sequences in a given query set. A multiple sequence alignment is represented in the form of a matrix with each DNA sequence occupying a row so that each nucleotide is placed in an appropriate column. Gaps are inserted between the nucleotide bases or between amino acid residues so that identical or similar characters are aligned in successive columns. In order to measure the degree of relatedness between sequences, weights are assigned to the aligned elements of sequences.

Traditionally, barcode regions were designed by first generating a multi-species alignment, and then, manually identifying conserved regions in that alignment and finally an algorithm was used to estimate the melting temperature of candidate primer sequences within the conserved regions. Till now, the most popular methods to find out conserved regions, start with a given multiple sequence alignment. One of them is a window-based approach. In this method a window of fixed length is moved down the sequence alignment and the content statistics are calculated at each position where the window is moved to (Nekrutenko and Li, 2000, Rice *et al.*, 2000). Since an optimal region could span several windows, the window-based approach suffers from the limitation of sometime failing in finding the exact locations of some interesting regions (Lin *et al.*, 2002).

Stojanovic *et al.* (1999) has proposed some more algorithms. The simplest of these is to compute level of similarity in each column of alignment matrix and find blocks that fit user defined threshold for degree of similarity per column and the length of block. For example column agreement 50%, and minimum length 5 as shown in figure 1.1. This approach known as column agreement approach, however, does not take into account the affect of nucleotide frequency. Schneider *et al.* (1986) proposed an optimization in the method that takes into account other informations like nucleotide similarity and overall nucleotide composition in the form of a score that measures its information content. Stojanovic *et al.* (1999) proposed three more methods for finding conserved regions from multiple alignment. These methods are all based on a columns score that depends either on the evolutionary relationships among the sequences implied by a given phylogenetic tree, or based on the longest region in which no row differs from a specified "center" sequence in more than *k* positions or based on the longest region in which no row differs from an unknown "center" sequence in more than *k* positions.

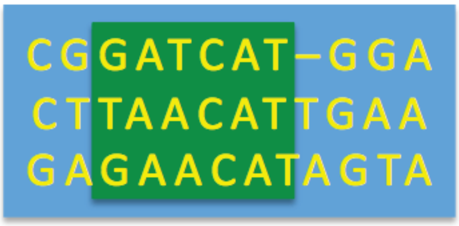


Figure 1.1: An example of 50% column agreement and and minimum length 5.

However, the algorithms of Stojanovic *et al.* (1999) suffer from a drawback that they can erroneously report the entire alignment as a single conserved region. This is because these methods are based on assigning a numerical score to each column of a multiple alignment and then looking for column's lengths with high cumulative scores. Since the assigned scores may be all positive (*e.g.* in the information content case), each examined column could increase the cumulative score and hence the entire alignment could be reported erroneously as a conserved region. To overcome this problem the author proposed that a positive anchor value should be subtracted from the column score. However, determining such an anchor value appropriately for each dataset could make the use of these methods very complicated.

Some more algorithms have been proposed to find conserved regions from protein sequence alignment. Livingstone and Barton (1993) proposed a method in which sequences in the alignment are gathered into subgroups on the basis of sequence similarity, functional, evolutionary or other criteria. All pairs of subgroups are then compared to highlight positions that confer the unique features of each subgroup. The algorithm is encoded in the computer program AMAS[5] (Analysis of Multiply Aligned Sequences). This algorithm was used in the alignment of 67 $SH2$ domains where patterns of conserved hydrophobic residues that constitute the protein core were highlighted. Although doing multiple alignment to locate the conserved regions within DNA or protein sequences seems the most straightforward solution but actually it is not an efficient solution, primarily because multiple alignment is a complex problem itself. Although there exist many efficient algorithms for achieving multiple alignment like dynamic programing but they are not efficient enough for aligning fully sequenced whole genomes of several giga bytes. So there is a strong need to look for more elegant solutions to scan the input sequences and find out conserved regions.

1.6.3 Finding Conserved Region Without Multiple Alignment

There are some algorithms available for finding conserved regions which are not based on the processing of sequence alignment. These algorithms can be divided into two main types, either combinatorial or probabilistic. Combinatorial algorithms are devised to tackle with combinatorial problems which involve the study of the number of ways of selecting or arranging objects from a finite set. Searching for patterns in a given data is a common example of combinatorial problems.

Combinatorial Methods

Suffix Tree: In computer science, suffix tree is a data structure widely used for searching a pattern in a string. The principal concept of suffix tree is that any string of a specified length can be broken down into suffixes (a string of length n has n suffixes), and these suffixes can be stored in a tree which allows fast and easy implementation of many string operations. A suffix of a string is a subset of symbols placed after the stem of the string,

[5] (`http://www.compbio.dundee.ac.uk/manuals/amas/`)

where the order of the elements is preserved. The suffix tree for the string τ over the alphabet σ is a tree with a set of nodes and edges such that:

- The tree starts from the root node and nodes are connected by edges.
- Each edge is labeled with a non-empty word and no two edges outgoing from the same node are labeled with the same word.
- Edges spell non-empty strings.
- The concatenation of the labels of the path from the root to a leaf spells one of the suffix of τ and the tree has n leaves.

Creating such a suffix tree and searching a pattern in it are both linear in time. Creating requires time $O(n)$ and searching requires time $O(m)$, where m is the length of the pattern. Since there is a path from root of the tree to each suffix of the string, hence at most m comparisons are needed to find a pattern of length m. This $O(m)$ time complexity is already good because a sequential search requires $O(n)$ time. The total length of all the strings on all of the edges in the tree is $O(n^2)$ but each edge can be stored as the position and length of a substring of τ, giving a total space usage of $O(n)$ computer words. The worst-case space usage of a suffix tree is seen with a fibonacci word, giving the full $2n$ nodes.

Suffix trees have been extensively studied and widely used. The first linear-time suffix tree algorithm was developed by Weiner (1973). A more space efficient algorithm was produced by McCreight (1976), and Ukkonen (1995) developed an "on-line" variant of suffix tree. The important bioinformatics applications of suffix trees include, finding the longest repeated substring (Weiner, 1973), computing substring statistics (Apostolico and Preparata, 1985), string comparison (Ehrenfeucht and Haussler, 1988), approximate string matching(Landau and Vishkin, 1989), identification of sequence repeats (Kurtz and Schleiermacher, 1999), multiple genome alignment (Hohl *et al.*, 2002) and selection of signature oligonucleotides for DNA arrays (Kaderali and Schliep, 2002).

Although the linear time and space complexity of suffix trees is quite attractive for many biological applications, this bound is not good enough for very large problems. Especially the space complexity is big hinderance for storing large amount of data and searching in it. This led to the development of structures such as Suffix Arrays to conserve memory.

Suffix Array Suffix arrays were introduced by Manber and Myers (1990) to find the strict repeats in a string. They are very space efficient and use almost three to five times less space than suffix trees. Suffix array is a sorted list of all suffixes of string τ. Formally it is defined as an array of the integers in the range 1 to n, and gives the positions of suffixes of τ in lexicographic order. Suffix array stores only positions of suffixes and does not store any other information about the alphabet of τ, therefore the space required by suffix array is modest and it can be stored in n computer words.

Inserting the suffixes into the array requires $O(n^2)$ time if suffixes are inserted one by one and making sure that new inserted suffix is in its correct place. However, using some efficient string sorting algorithms like the ones developed by Baer and Lin (1989), array construction can be done in $O(n \log n)$ time. Space complexity of suffix tree is linear and is $O(n)$ because for each suffix we need one position in array. Manber and Myers (1990) presented an algorithm which uses two arrays instead of one; the first to store positions of suffixes and the second to store the information about least common prefix (lcp) of adjacent elements in suffix array. Construction of suffix array and its lcp information requires $O(n \log n)$ time and $O(3n)$ space in the worst case. Although $2n$ are occupied by suffix array and lcp array, another n integers are required during their construction. Suffix arrays can be used instead of suffix trees in many applications, especially they are very efficient for large datasets. An algorithm developed by Ko and Aluru (2003) derived from suffix tree construction algorithm of Farach (1997) reduces the array construction time from $O(n \log n)$ to $O(n)$.

The suffix array of a string can be used as an index to quickly locate every occurrence of a substring. According to this algorithm, using these two data structures and applying a simple binary search requires $O(m + [\log_2(n-1)])$ time and $O(2n)$ space.

Karp-Miller-Rosenberg (KMR) Algorithm The *KMR* (Karp *et al.*, 1972) is an old algorithm and was the first almost linear algorithm that finds repeated identical patterns in three structures: strings, arrays and trees. Although the algorithm is similar for other structures but we will only concentrate on string structures where the pattern to be found is a substring. For a given string τ of length n defined over the alphabet σ, *KMR* can find:

- The longest repeated substring
- All k-length repeated substrings
- The positions of each instance of repeated substring

KMR is based on the idea of a family of equivalence relations on the set of positions of the input string, denoted by $\{i \; E_k \; j\}$, where i and j are two positions in τ and $i, j \in \{1, 2, ..., n-k+1\}$. The two positions i and j are called $k - equivalent$ if the words of length k starting at these positions in τ are identical. The algorithm iteratively constructs the equivalence relation over the string τ starting from $k = 1$ and doubling the size of k at each iteration. To find a pattern of length d, successively construct relations $E_2, E_4, E_8, \ldots, E_r$ where $r = 2^{\lfloor \log_2 d \rfloor}$. If d is a power of 2 then end otherwise $d < 2r$ and $\{i\ E_{r+(d-r)}\ j\}$ gives the required solution.

With this technique the algorithm progressively constructs all occurrences of $2k$-length repeated patterns starting from all k-length repeated patterns. Such a construction of equivalence relation requires at most $\log k$ steps and hence the time complexity of KMR is $O(n \log k)$. One important space saving strategy used in KMR is that at each position in τ, a number called class code is used to identify the pattern starting at that position. Identical patterns of a certain length have same class codes. With this technique KMR has a space complexity of $O(n)$.

One variant of the KMR algorithm called KMR clique or KMRc given by Soldano *et al.* (1995) can be used that allows to tolerate some errors thus finding approximate repeats.

Matinez's Sorting Algorithm The algorithm of Martinez (1983) displays a priori unknown identically repeated patterns in several molecular sequences. The algorithm solves the problem of finding repeated patterns as a recursive sorting problem. The method used to find repeats is quite similar to the algorithm of KMR. For a string τ defined of the alphabet σ where the $|\sigma| = m$, we construct a sequence P of pointers such that pointer value $P[i]$ is the location of the i^{th} element in τ. The next step is sorting of P so that it constitutes an ordering of τ. That is, $P[i] < P[j]$ or $P[i] > P[j]$ or $P[i] = P[j]$ according to whether $\tau[P[i]] < \tau[P[j]]$ or $\tau[P[i]] > \tau[P[j]]$ or $\tau[P[i]] = \tau[P[j]]$ respectively. Such a sorting of P results in grouping of same kind of elements of τ and at most m groups are possible. In the next iteration each group of P is again sorted such that in the resulting subgroups two pointer values point to the same one if and only if the elements immediately following the ones they point to, are equal. The process continues when no subgroups contain more than one pointer value and the final result is an ordering of P. With such repeats finding strategy the algorithm has a time complexity of $O(n \log n)$, however the overall speed of algorithm depends on the sorting algorithm employed. Space complexity of this algorithm is $O(n)$, no significant storage space is required beyond that necessary for the string τ and its pointer sequence P.

Probabilistic Methods

GIBBS Sampling Gibbs sampling is a statistical approach for finding strict and approximate repeats. It is a sampling algorithm which can generate a sequence of samples from a given distribution of two or more random variables. The purpose of these samples is to approximate the joint distribution, marginal distribution or to compute the integral of some function. Gibbs sampling exploits the idea of sampling from a conditional distribution. In this sampling technique, we sample each variable separately from a conditional probability where all the other variables are taken as fixed using the latest values of these variables in each step. For example, in order to take k samples from a joint distribution $P(x_1, x_2, \ldots, x_n)$, where i^{th} sample is represented as $X^i = \{x_1^i, x_2^i, \ldots, x_n^i\}$, we start with first known sample X^0 and generate the next sample set by sampling each variable x_j^i from distribution $P(x_j^i | x_1^i, \ldots, x_{j-1}^i, x_{j+1}^{i-1}, \ldots, x_n^{i-1})$.

Lawrence *et al.* (1993) have used Gibbs sampling technique for multiple alignment to detect subtle sequence signals. To find mutually similar segments of width w in given n strings $(\tau_1, \tau_2, \ldots, \tau_N)$ they construct two evolving data structures. The first one, called probabilistic model of pattern description, describes the probability of occurrence of each symbol on each position of the pattern along with probabilistic model of background frequencies with which residues occur in sites not described by the pattern. The second data structure simply keeps the starting position of the pattern in all sequences. Best pattern is obtained by locating the alignment that maximizes the ratio of pattern probability to the background probability. Starting from randomly chosen initial positions of the pattern in all sequences, the algorithm in first step, updates the pattern probabilities for all but one sequence chosen either randomly or in a certain order. Then, in second step, the algorithm finds a new random position of the pattern in the sequence ignored in first step, using random sampling from probabilistically weighted all possible segments of this sequence using pattern and background probabilities. The main motivation for this work comes from the high dimensionality of the search space and the existence of many local optima. Due to stochastic sampling, the algorithm does not get stuck in local optima. Moreover, the large search space is explored one dimension at a time.

Space complexity of this approach is $O(n)$ where major part of memory is used to save input sequences whereas memory used by three data structures is quite negligible. Time complexity of this algorithm is $O(tnlw)$ where t is the number of times algorithm executes before convergence and l is average length of all the input sequences. If the common pattern exists at roughly equal probability in the input sequences then time complexity tends to be linear, i.e. $O(n)$. Gibbs sampling is both fast and sensitive, but because it is a stochastic method, it may not sample all search space and may produce slightly different results at each run.

1.6.4 Some Important Properties Related To Oligonucleotide Primers

DNA barcoding makes use of a short DNA sequence (with some interesting properties) to identify it as belonging to a particular taxa. We have already stated that an ideal barcode marker consists of two conserved regions flanking a central variable region. We have also given some details on methods to identify conserved regions from genomes. In actual experiments these conserved regions are used as oligonucleotide primers or probes, where the function of these primers or probes is to detect the target barcode region from an organism's DNA. These oligonucleotides are called primers when used in PCR experiment and they are called probes when used for hybridization. Hybridization is the formation of specific double stranded nucleic acid molecules from two complementary single stranded molecules.

For a PCR, template DNA and primers are mixed together with other reactants (*e.g.* nucleotides, DNA polymerase *etc*). During each PCR cycle, the double stranded DNA is melted into two single strands. The primer pairs added in the mix can hybridize at loci flanking the region that we want to amplify (annealing step). Then the DNA polymerase can extend those primers, thus building a new double stranded DNA corresponding to the selected region (extension steps). Multiple repetitions of such a cycle lead to the over-production of the selected region (Saiki *et al.*, 1985). In order to achieve the strong association between primers and target sequences, certain properties of primers need to be considered. They include, for example, self-complementarity, annealing and extension temperature and length of primer sequences. Self-complementarity is the phenomena when a primer contains some nucleotide bases which are complement of each other, or one primer contains a sequence which is complement of some other primer sequence in the mixture. In both of these cases primers can self-complement making a double strand among themselves (primer dimers) and primers are not available to hybridize to the target sequence (Burpo, 2001). Formation of stable primer-target duplexes requires low self-complementarity.

It is also important that melting temperature (T_m) of both primers is similar to ensure as much consistent performance as possible between forward and reverse primer pairs. This is because the actual hybridization temperature determines the outcome of the experiment. For PCR primers normally the T_m of 54 °C or higher is preferred. A third condition that needs to be considered is 3′ strict match. Primer and template sequences may allow some mismatches but it is preferred that these mismatches do not occur at 3′ end. According to Kwok *et al.* (1990) extension of primer and target sequence is most dependent on outermost 3′ base pairing, less on 2^{nd} and 3^{rd} last pairs and even less on the other pairs. This implies that 3′ end of primer and target sequence should match perfectly. The last important parameter to be considered is primer length. Primer length is a function of universality, hybridization stability, and cost-minimization by seeking shortest possible oligonucleotides. A length of $18 - 22$ bp is considered as the suitable length. This length is long enough to allow specificity and short enough to allow cost minimization.

These were some considerations in the terms of primers design for PCR applications. In the next section I will describe some programs available for designing primer pairs and barcode markers. Almost all of the programs give enough weight to the criteria listed above.

1.6.5 Barcode Designing Tools

Based on the algorithms described above for finding conserved regions and keeping in consideration the primer properties, several barcode designing tools exist today. However it will be more appropriate to use the term primer design instead of barcode design because these tools concentrate more on primer design and the thermodynamic properties of primers pairs and give no measures about the discrimination capacity of the region amplified by the primer pairs. Most of the tools that exist today are easily usable in the context of *sensu stricto* barcoding when we want to adapt standard barcode primers to a new clade. But they are less adapted for metabarcoding or environmental barcoding. In this section, we will have a detailed look on most of the important tools that have been designed for this purpose.

The first program inline is perhaps *PRIMER* (*Primer 0.5*) developed by Whitehead Institute/MIT Center for Genome Research but this program was never published. A complete rewritten version of *PRIMER* (*Primer 0.5*) exists in the form of *Primer*3 (Rozen and Skaletsky, 2000). It takes as input a single sequence and selects single primers or PCR primer pairs considering oligonucleotide melting temperature, length, GC content, primer-dimer possibilities, PCR product size and positional constraints within the source sequence. *Primer*3 also provides some objective functions to be computed for each primer pair. They include: checking each primer pair against a mispriming library (which means that primer pairs should not amplify any of the non-target sequences specified in the mispriming library) and checking the primers for self-complementarity. Nevertheless the computation of objective functions increase the running time of the program. The most time expensive operation is to check each primer pair against a mispriming library. In this case Primer3 adopts a very rigorous approach of locally aligning each candidate primer against each library sequence and rejecting those primers for which the local alignment score exceeds a specified weight. Running time of *Primer*3 is also dependent on size of input sequence and hence it is a linear function of sequence size. Primer3 is perhaps an ideal solution as it provides a lot of adjustable parameters and it is the most widely used program but since it seeks to amplify a single target sequence, it cannot be used to design universal primers amplifying a large number of target sequences. Moreover the overhead of alignment for excluding the non-target sequences amplification makes this program infeasible for large applications.

UniPrime (Bekaert and Teeling, 2008), *QPRIMER* (Kim and Lee, 2007), and *Primaclade* (Gadberry *et al.*, 2005) are three programs based on the alignment of multiple sequences to find the primer pairs.

UniPrime takes *Genbank* GenID of the target locus as input and selects the prototype sequence (mRNA sequence of longest isoform of gene). This prototype sequence is then used as a query sequence in *Blastn* search to search for all highly similar homologous sequences. Stored sequences are concatenated into a single file and then aligned using $TCoffee$ program (Notredame *et al.*, 2000). From this alignment a consensus sequences is inferred and all possible primers along the consensus sequence are generated by $Primer3$.

QPRIMER is a web-based application that designs conserved PCR and RT-PCR primers from multiple genome alignment making use of a genome browser (*Pygr*) and $Primer3$ programs. *Pygr*[6] (Python Graph Database Framework for Bioinformatics) is an open source program that allows sequence and comparative genomics analyses. It can query large sequence databases or multiple genome data sets to find regions of interests. *QPRIMER* supports human, mouse, rat, chicken, dog, zebrafish and fruit fly sequences to design primers. This program allows its users to browse a specific gene of interest using genome browser based on genomic location. Users can select any region in the gene structure as a target for amplification. For the selected gene region, *QPRIMER* uses *Pygr* to extract the sequence from multiple alignment dataset. It then uses $Primer3$ program to design primer pairs from the extracted data set. *QPRIMER* selects primers from only exonic regions. The major disadvantage of such primer design approach is that primers are selected only from a single sequence and non-target sequences are not allowed. Moreover this type of application can only be useful for selecting primer pairs from standard genes which are known to be conserved.

Primaclade is also a web-based application and is based on multiple genome alignment to infer conserved regions. It accepts a multiple species nucleotide alignment file saved as Clustal (Thompson *et al.*, 1997), EMBOSS (Rice *et al.*, 2000) or any other alignment format as input and identifies a set of degenerate PCR primers that will bind across the alignment. To select the primer pairs, *Primaclade* computes a consensus sequence

[6] (`http://bioinfo.mbi.ucla.edu/pygr/`)

from the alignment file. It then splits the alignment file into individual sequences and uses *Primer*3 program to compute a set of exhaustive primer pairs for each individual sequence from alignment file. To compute a large number of Primers, *Primaclade* runs *Primer*3 program eleven times for each sequence starting from primer length of 18 bp and increasing the length by 1 bp each time eventually terminating at primer length of 28 bp. After generating primer pairs for each sequence from the alignment file, it compares them to the corresponding nucleotides in consensus sequence to see if consensus sequence contains the correct number or fewer degenerate nucleotides. In this case primer is saved otherwise the pair is discarded.

This approach of primer design based on alignment is although effective, it provides a limited number of barcode markers, because only those conserved regions are identified which are specific to a particular gene.

Some programs have been designed in the context of environmental applications like *Greene SCPrimer* (Jabado *et al.*, 2006) and *PrimerHunter* (Duitama *et al.*, 2009). Both of these programs have been designed for PCR detection of viruses which are *sensu lato* barcoding applications. *Greene SCPrimer* is also based on the processing of a multiple sequence alignment. It determines the optimum primer pairs from a nucleic acid sequence alignment by first constructing a phylogenetic tree to identify candidate primers and then using a greedy algorithm to identify minimum set of primers that amplifies all members of alignment. The exact algorithm is as follows. From a multiple alignment of sequences, sub-alignments of length appropriate for PCR primers are extracted and only unique strings are kept for further processing. For the short sequences that are kept, a similarity matrix is generated using pairwise alignment. This similarity matrix is used to generate a phylogenetic tree using a hierarchical clustering algorithm based on Euclidean distance using an open source clustering library (de Hoon *et al.*, 2004). At each node of the phylogenetic tree, a consensus sequence is computed and then primers are checked and filtered for physical constraints like T_m, *GC* content and degeneracy *etc*. After scoring the primers, greedy algorithm is performed to keep only a minimum number of primers which amplify all the sequences in the alignment and the last step is to identify primer pairs with matching T_m suitable for amplifying products of a specific size range. The time complexity of tree building step is $O(n^3)$, primer scoring step is linear in time, however the third step of primer minimizing has complexity of $O(n \log n)$. The last step of building primer pairs is linear in time.

The other program for designing PCR primers for viruses is *PrimerHunter* (Duitama *et al.*, 2009). This program has been designed to select highly sensitive and specific primers for virus sub types. The tool takes as input two fasta files, one containing the target sequences and the other containing non-target sequences. Primers are selected such that they efficiently amplify any one of the target sequence and none of the non-target sequences. The program uses some hash tables to build primers making sure that 3′ end does not allow mismatches. *Primer Hunter* uses nearest neighbor thermodynamics model of SantaLucia and Hicks (2004) for calculating accurate melting temperature.

Although both *Greene SCPrimer* and *PrimerHunter* are *sensu lato* barcoding applications, the efficiency of both programs is a big question mark. *Greene SCPrimer* is based on the processing of alignment and constructs phylogenetic tree which are very expensive computations in time and hence for large sequences this program cannot be efficiently used. *PrimerHunter* is based on thermodynamics model and it also needs to do a lot of computation to see if a primer pair should be selected or not, and hence again for small sequences this program is good but for larger sequence databases, it is not efficient enough. A comparison of the main features of some important existing primer and probe selection tools is given in (Duitama *et al.*, 2009).

1.6.6 Our Contribution

Our work builds on the idea that primer design is an optimization problem that can be solved by adapting methods from computer science. Using sequence alignment to locate conserved regions is a time consuming method and is not efficient enough to be used for locating conserved sequences from whole genomes of several hundred thousand base pairs. Hence this method is only suitable for well-known sets of genes. We have seen that most of the programs make use of *Primer*3 to select primers, but this approach seeks to amplify a single target sequence and does not guarantee amplification sensitivity in the presence of high sequence heterogeneity as in the case of environmental samples, where DNA from different species is present in the mixture. Almost all of the programs focus on the selection of best primer pairs by providing a lot of adjustable parameters but no program considers the importance of whole genome scanning for identifying the universal primer pairs. Moreover no program gives any indication about the quality of primers and barcode regions in terms of their amplification and taxon discrimination capacity. In this context we have developed two quality measures and a program called *ecoPrimers* keeping in consideration the efficiency in terms of time and memory to be able to scan large databases of long genomes and the missing key features in already existing programs. This result has been described in chapter 2 and 3 of this thesis.

1.7 More Deep Into DNA Metabarcoding

Metabarcoding or environmental barcoding requires short barcode markers because DNA is mostly degraded in environmental samples. These short barcode markers may not have high discrimination capacity and hence a single barcode marker may not identify all of the organisms from an environmental sample. In such a case we can imagine that species identification could be carried out by the combined analysis of several short universal barcode markers. By using several short markers in combination, the total number of identified taxa can be increased. In this technique the important problem is the selection of the best set of markers from a pool of several barcode markers to achieve maximum number of identifications, keeping at the same time the size of set to be as minimum as possible. Such a problem is a type of combinatorial problems and more precisely it is a set cover problem that has been proven to be NP-complete (Lund and Yannakakis, 1994). The class of NP-complete problems has the important property that no polynomial time algorithm for any of its members exists to date and in case a polynomial time algorithm for one NP-complete problem was found, all could be solved in polynomial time. They are therefore considered as inherently intractable from a computational point of view. Thus, in the worst case any algorithm that tries to solve an NP-complete problem requires exponential run time. In order to efficiently deal with NP-complete problems there are several metaheuristic approaches available, which can be used to find the near optimal solution. I will discuss the details of our sets approach in chapter 3, however, I will give a brief introduction to combinatorial problems and metaheuristic approaches in this section.

1.7.1 Combinatorial Problems And Approximate Methods

Many of the problems in the field of bioinformatics correspond to hard combinatorial problems. The field of combinatorics deals with the study of the number of ways of selecting or arranging objects from a finite set or possibly countably infinite set. The object may be a subset from a large given set, an integer number, a subgraph or a permutation. The finite set or countable infinite set is called the solution space. Blum and Roli (2003) formalized such a problem as:

A combinatorial problem $\mathcal{P} = (\mathcal{S}, f)$ can be defined by

- a set of variables $X = \{ x_1, x_2, x_3, ...x_n \}$
- a variable domain $D_1, D_2, ...D_n$

- and an objective function f to be maximized, where $f : D_1 \times D_2 \times ...... \times D_n$ and $D \in R$

The set of all possible assignments

$$\mathcal{S} = \{ \text{s} = \{ (x_1, v_1), (x_2, v_2), ... (x_n, v_n) \} \quad | \quad v_i \in D_i \}$$

is called a search or solution space, as each element of the set can be seen as a candidate solution. To solve such a combinatorial problem we need to find a solution $s^* \in \mathcal{S}$ such that the value of objective function f is maximized that is $f(s^*) \geq f(s) \, \forall \, s \in S$. s^* is called a globally optimal solution of $(\mathcal{S}, \text{f})$.

The algorithmic approaches to such combinatorial problems can be classified as either exact or approximate. Exact algorithms are guaranteed to find an optimal solution in finite time by systematically searching the solution space. For example, for the above sets problem, the most straightforward exact solution is to simply enumerate the full solution space and choose the best set which maximizes the objective function. Yet such an algorithm is infeasible because the search space of candidate solutions grows exponentially as the size of the problem increases. To practically solve these problems, one often needs finding good, approximately optimal solutions in reasonable time, that is, polynomial time. Approximate algorithms cannot guarantee optimality of the solutions they return; the essence of an approximate method is to find the good solution in a significantly reduced amount of time and this is exactly what is required in many problems related to molecular biology and bioinformatics. In the field of bioinformatics, researchers rarely need an optimal solution, in-fact people want robust, fast and near-optimal solutions. In this context, the use of approximate methods provides an efficient and simple way of solving combinatorial problems.

Approximate methods can be divided into two different types; constructive methods and local search methods. Constructive algorithms generate solutions from scratch. They add components to an initial empty partial solution, until the solution is complete. They are typically the fastest approximate methods, but often return solutions of inferior quality when compared to local search algorithms. Local search algorithms start from some initial solution and try to find a better solution in an appropriately defined neighborhood of the current solution. In case a better solution is found, it replaces the current solution and the local search is continued from there. The neighborhood can be formally defined as a function $\mathcal{N}$:$\mathcal{S} \mapsto 2^{\mathcal{S}}$ that assigns to every $s \in \mathcal{S}$ a set of neighbors $\mathcal{N}(s) \subseteq \mathcal{S}$. $\mathcal{N}(s)$ is formally called the neighborhood of s. The initial solution could be any random solution or a well thought solution depending upon the problem. The most basic local search algorithm, called iterative improvement, repeatedly applies these steps until no better solution can be found in the neighborhood of the current solution and stops in a local optimum. The main disadvantage of this algorithm is that it may stop at poor quality local minima, where a local minima can be defined as:

Definition 4. A local minima or a local minimum solution with respect to a neighborhood structure $\mathcal{N}$ is a solution s' such that $\forall\, s \in \mathcal{N}(s') : f(s') \leq f(s)$. s' is called a strict locally minimal solution if $f(s') < f(s)\ \forall\, s \in \mathcal{N}(s')$.

One possibility to improve the performance of local search algorithm could be to increase the size of the neighborhood used in the local search algorithm. With this strategy, there is a higher chance to find an improved solution, but it also takes longer time to evaluate the neighboring solutions, making this approach infeasible for larger neighborhoods. One more possibility could be to restart the local search algorithm multiple times, each time starting from a new randomly generated solution until some stopping criterion. The best local minimum found during this approach could be accepted as the final solution. While, this approach may give good results for small data sets, for increasing problem size, it could become infeasible to run the local search algorithm many times. In order to avoid the problem of trapping in local minima, some extensions of the local search algorithms have been proposed. The techniques to improve local search algorithms by avoiding the problem of local minima are called metaheuristics.

1.7.2 Metaheuristics

One of the emerging class of approximate methods is metaheuristics that has been designed to solve a very general class of combinatorial optimization problems. The term metaheuristics was first introduced by Glover (1986) however Kirkpatrick *et al.* (1983) had already proposed a well know metaheuristic technique called Simulated Annealing (SA) in 1983. According to Glover "a metaheuristics refers to a master strategy that guides and modifies other heuristics to produce solutions beyond those that are normally generated in a quest for local optimality". Metaheuristics are not problem specific, they provide a general algorithmic framework which can be applied to different optimization problems with relatively few modifications and using domain specific knowledge to make them adapted to a specific problem (Blum and Roli, 2003). There is no standard and commonly accepted definition for the term metaheuristics, however, in the last few years different researchers tried to propose different definitions for the term (Osman and Laporte, 1996, Stützle, 1999, Voss *et al.*, 1999). The simplest of these definitions is one given by Osman and Laporte (1996), which says:

Definition 5. ”A metaheuristic is formally defined as an iterative generation process which guides a subordinate heuristic by combining intelligently different concepts for exploring and exploiting the search space, learning strategies are used to structure information in order to find efficiently near-optimal solutions.”

The main goal of metaheuristics algorithms is to avoid the disadvantage of iterative local search to escape from local minima. Different strategies have been devised to achieve this. They include either allowing the low quality solutions or generating new starting solutions in a more intelligent way than just using random initial solutions. Many of the proposed methods make use of objective functions, information of previously made decisions or probabilistic models during the search to escape from local minima.

Metaheuristics methods have many interesting applications in almost all fields of scientific research including psychology, biology and physics. A number of applications have been discussed (Beer, 1996, Osman and Kelly, 1996, Vidal, 1993) and a useful metaheuristic survey is given (Osman and Laporte, 1996). In this section we will talk about the two most studied and used methods called Simulated Annealing and Tabu Search. We make use of these methods in our primers sets approach which is the chapter 3 of this thesis.

Simulated Annealing

Simulated Annealing (SA) is the oldest among the metaheuristics and was independently proposed by Kirkpatrick *et al.* (1983) and Černý (1985). The concept of simulated annealing algorithm is taken from physical annealing in metallurgy. The technique of physical annealing involves heating and controlled cooling of a material to increase the size of its crystals and reduce their defects. Controlled cooling means to lower the temperature very slowly and spending a long time at low temperatures in order to grow solids with a perfect smooth structure. If cooling is done too fast, the resulting crystals will have irregularities and defects. This undesirable situation is avoided by a careful annealing process in which the temperature descends slowly through several temperature levels and each temperature is held long enough to allow the solid to reach thermal equilibrium. Such a state corresponds to a state of minimum energy and the solid is said to be in a ground state. There exists a strong analogy between combinatorial optimization problems and physical annealing of solids (crystals), where the set of solutions of the problem can be associated with the states of the physical system, the objective function corresponds to physical energy of the solid, and globally optimal solution corresponds to the ground state of solids.

Simulated annealing uses the idea of basic local search however it allows moves of inferior

```
$s \leftarrow GenerateInitialSolution()$
$T \leftarrow T_0$
while termination conditions not met do
  $s' \leftarrow PickAtRandom(\mathcal{N}(s))$
  if $(f(s') < f(s))$ then
    $s' \leftarrow s$
  else
    Accept s' as new solution with probability $P_{accept}(T, s, s')$
  end if
  $Update(T)$
end while
```

Figure 1.2: Simulated Annealing Algorithm

quality (Reeves, 1995) to escape from local minima. The algorithm starts by generating a tentative solution s' and initializing a temperature T. This solution s' is accepted if it improves the objective function value, however, if s' is worse than the current solution, it is accepted with a probability which depends on the difference $\triangle = f(s) - f(s')$ of objective function for current solution s, the tentative solution s' and temperature T. The probability of acceptance is computed by following Boltzmann distribution as $e^{-\triangle/T}$. The probability P_{accept} to accept worse solutions is defined as:

$$P_{accept}(T, s, s') = \begin{cases} 1 & \text{if f(s) < f(s')} \\ e^{-\triangle/T} & \text{otherwise} \end{cases}$$

A simpler version of algorithm for simulated annealing is shown in figure 1.2.

At the start of algorithm, the temperature is high and the probability to accept inferior quality solutions is also high but it decreases gradually, converging to a simple iterative improvement algorithm when the temperature is lowered gradually. In the beginning when the probability to accept inferior quality solutions is high, the improvement in the final solution is low and a large part of solution space is explored however the algorithm eventually tends to converge to local minima when the probability is lowered. The probability of accepting inferior quality solutions is controlled by two factors: the difference of the objective functions and the temperature. It means that at fixed temperature, the higher the difference $\triangle = f(s) - f(s')$, the lower the probability to accept a move from s to s'. On the other hand, the higher the temperature, the higher the probability of accepting inferior quality solutions.

An appropriate temperature lowering system (defined by $Update(T)$ function in figure 1.2) is crucial for the performance of algorithm. Such a system is called *annealing schedule* or *cooling schedule*. It is defined by an initial temperature T_0 and a scheme saying how the new temperature is obtained from the previous one. Such a system also defines the number of iterations to be performed at each temperature and a termination condition. An appropriate *cooling schedule* guarantees the convergence to a global optimum, however such a schedule is not feasible in applications because it is too slow for practical purposes Therefore, faster cooling schedules are adopted in applications. One of the most used cooling schedule follows a geometric law: $T_{k+1} = \alpha T_k$ where $\alpha \in (0, 1)$ and k is the number of iterations (Blum and Roli, 2003). Such a schedule corresponds to an exponential decay of the temperature. More successful variants are non-monotonic cooling schedules (Lourenço *et al.*, 2001), which are characterized by alternating phases of cooling and reheating, thus providing a balance between revisiting some regions and exploring the new regions of search space. However in actual applications one good strategy could be to vary the cooling rules during the search, like temperature could be constant or linearly decreasing at the beginning in order to sample the search space and then T might follow a geometric rule at the end of search to converge to a local minimum. SA has been successfully applied to several combinatorial optimization problems, such as the Quadratic Assignment Problem (Connolly, 1990) and Job Shop Scheduling Problems (van Laarhoven *et al.*, 1992).

Tabu Search

The basic idea of Tabu Search (TS) was first introduced by Glover (1986). This is among the most cited and used metaheuristics for combinatorial optimization problems. The basic idea of TS is to use information about the search history to guide local search approaches to escape from local minima and to implement an explorative strategy. This is done by using a short term memory called *tabu list*, which is a small list for storing some forbidden solutions.

TS uses a local search algorithm that in each step tries to make the best possible move from current solution s to a neighboring solution s' even if that move gives an inferior quality value of objective function. To prevent the local search to immediately return to a previously visited solution and to avoid cycling, moves to recently visited solutions are forbidden. This can be done by keeping track of previously visited solutions by adding those solutions to *tabu list* and forbidding moving to those. These moves are forbidden for a pre-specified number of algorithm iterations for example t iterations.

Forbidding possible moves dynamically restricts the neighborhood $\mathcal{N}(s)$ of the current solution s to a subset $\mathcal{A}(s)$ of admissible solutions. At each iteration the best solution from the allowed subset $\mathcal{A}(s)$ is chosen as the new current solution. Additionally, this solution is added to the *tabu list* and one of the solutions that were already in the *tabu list* is removed usually in a FIFO (First In First Out) order. Basic algorithm for TS is shown in

```
$s \leftarrow GenerateInitialSolution()$
$s_{best} \leftarrow s$
$tabulist \leftarrow \emptyset$
while termination conditions not met do
    $\mathcal{A}(s) \leftarrow GenerateAdmissibleSolutions(s)$
    $s \leftarrow ChooseBestOf(\mathcal{N}(s) \mid tabulist)$
    $Update(tabulist)$
    if $(f(s_{best}) < f(s))$ then
        $s_{best} \leftarrow s$
    end if
end while
```

Figure 1.3: Simple Tabu Search Algorithm

figure 1.3. Removal of elements from *tabu list* is important because of two reasons. First, size of *tabu list* is kept small for fast access, so when the list is full and there is no more room for new elements, the one previously added elements has to be removed. Second it is important to remove already added solutions to list so that they can be made available for next moves. The algorithm stops when a termination condition is met or if the allowed set is empty, *i.e.* all the solutions in $\mathcal{N}(s)$ are forbidden by the tabu list, however this rarely happens because usually the size of *tabu list* is very small as compared to the actual neighborhood size $|\mathcal{N}(s)|$.

The size of the *tabu list* (*tabu size*) controls the memory of the search process. With small *tabu size* the search will concentrate on small areas of the search space and a large *tabu size* forces the search process to explore larger regions, because it forbids revisiting a higher number of solutions. The *tabu size* can be varied during the search, leading to more robust algorithms. One of the example of dynamically changing size of *tabu list* is presented in (Battiti and Protasi, 1997), where the *tabu size* is increased if solutions are repeated, while it is decreased if there are no improvements.

Another important thing to be considered is that the short term memory used as *tabu list* does not actually contain the full solutions because managing a list of solutions is computationally very inefficient. Instead of adding the complete solutions to the list, some attributes to the solutions are stored as storing attributes is much more efficient than storing complete solutions. Because more than one attribute can be considered, a *tabu list* is built for each of them. The set of attributes and the corresponding *tabu list* define the tabu conditions which are used to filter the neighborhood $\mathcal{N}(s)$ of a solution s and generate the allowed set $\mathcal{A}(s)$. Although managing attributes is more efficient than managing full solutions, yet it may introduce a loss of information, as forbidding an attribute means assigning the tabu status to probably more than one solutions as more than one solutions can have same attributes. A major disadvantage of this phenomena is that an unvisited good quality solution can be excluded from the allowed set. To

```
$s \leftarrow GenerateInitialSolution()$
$Initializetabulists(tl_1, tl_2, .......tl_n)$
$k \leftarrow 0$
while termination conditions not met do
    $AllowedSet(s,k) \leftarrow \{s' \in (\mathcal{N}(s)$ | s does not violate a tabu condition,
                                        or it satisfies at least one aspiration condition$\}$
    $s \leftarrow ChooseBestOf(AllowedSet(s,k))$
    $UpdateTabuListsAndAspirationConditions()$
    $k \leftarrow k+1$
end while
```

Figure 1.4: Tabu Search Algorithm with aspiration condition

overcome this problem, aspiration criteria are defined which allow to include a solution in the allowed set $\mathcal{A}(s)$ even if it is forbidden by tabu conditions. Aspiration criteria define the aspiration conditions that are used to increase the size of allowed set $\mathcal{A}(s)$ by adding more elements in it during the search process. The most commonly used aspiration criterion selects solutions which are better than the current best one. *Tabu Search* algorithm with aspiration condition is shown in figure 1.4.

To date, TS appears to be one of the most successful metaheuristics. For many problems, TS implementations are among the algorithms giving the best tradeoff between solution quality and the computation time required (Nowicki and Smutnicki, 1996, Vaessens *et al.*, 1996). However, for the empirical success of this algorithm a very careful choice of parameter value adjustments and implementation data structures is required which includes managing *tabu size*, deciding the number of iteration *t* for algorithm and carefully choosing the aspiration criteria.

1.8 DNA Sequence Analysis

Due to next generation sequencing techniques and the availability of large public databases, today a large amount of sequence data is available for genomics research. The two new sequencing technologies *i.e.* 454 GS FLX/Roche and Solexa/Illumina system have been producing data at ultrahigh rates (Bentley, 2006). For example 454 Pyrosequencing with its newest chemistry termed "Titanium" can generate approximately 1×10^6 sequence reads in one run, with read lengths of ≥ 400 bases yielding up to 500 million base pairs (Mb) of sequence. Similarly Solexa system using its iterative, sequencing-by-synthesis process can generate 2×10^9 sequence reads in one run, with read lengths up to 100 bases. Public databases are also expanding at an exponential rate due to such large amount of data produced.

The tremendous amount of data produced by next generation sequencing techniques has greatly helped scientists in many ecological applications for instance in viral population dynamics (Wang *et al.*, 2010) or to characterize the phylogenetic diversity within microbial communities through amplification of 16S rRNA genes (Huber *et al.*, 2007). However one potential issue in this regard that cannot be ignored is the presence of noise in this data. The large number of reads obtainable mean that the absolute number of noisy reads is substantial (Quince *et al.*, 2011). Thus it is important to distinguish true sequence diversity in the sample from errors introduced by the experimental procedure. In microbial biodiversity estimation, sequences are clustered into Operational Taxonomic Units (OTUs) that represent the traditional taxa, and diversity is measured by estimating number of such OTUs in a community. In some of the early studies on pyrosequenced 16S rRNA genes like given by Sogin *et al.* (2006), a larger number of OTU's were observed, many of which had very low frequencies. These low frequency OTU's were considered as rare taxa and gave rise to the phenomena of ″rare biosphere″. In this same article Sogin *et al.* (2006) said that large number of reads produced in a single pyrosequencing run can provide unprecedented sampling depth and thus the rare biosphere is substantially larger and more diverse than previously appreciated. However, recent studies (Kunin *et al.*, 2010, Quince *et al.*, 2009) have shown that intrinsic error rate of pyrosequencing reads could lead to overestimates of the number of rare taxa and low frequency OTUs are actually generated by noise.

Occurrence of errors in sequence databases is also frequent, because nearly every time a listed gene is sequenced a second time, errors are reported. The incidence of corrections added to sequence data banks demonstrates that errors occur regularly (Clark and Whittam, 1992). The presence of errors in sequences can have adverse affects *e.g.* errors can cause non-polymorphic sites to appear polymorphic and vice versa. Moreover, errors can change one polymorphic site into a different polymorphic site by altering the frequency at which the 2 alleles appear in the sample (Johnson and Slatkin, 2008). Analyses of sequence variation in species with very low sequence diversity are particularly sensitive to such errors, because the signal-to-noise ratio is lower than that for species with relatively high levels of sequence diversity. Regardless of the source of errors, it is clear that presence of errors in sequences can have severe affects on molecular evolutionary analysis. In order to avoid making wrong conclusions it is important to be able to differentiate the erroneous reads from genuine sequences. For this purpose it becomes essential to understand different factors generating errors, learn the behavior of errors and present an error model based on the behavior under certain conditions.

In the context of this study, the term “sequence errors“ means the total number of erroneous nucleotides between an actual gene and the sequence as it appears in a data bank or generated by a sequencer. Total errors represent accumulation of mistakes generated by degradation of DNA as in the case of ancient DNA and environmental DNA, PCR induced point mutations and chimeras and errors generated during the process of sequencing due to sequencing chemistry. The errors generated in one step can pass to the next step making it more difficult to identify that at which step a particular error was generated. In this section, we will briefly talk about different types of errors produced due to environmental and experimental constraints.

1.8.1 Errors Due To DNA Degradation

Recent advances in molecular genetics have allowed DNA to be extracted, amplified and sequenced from ancient tissues. However, the validity of an ancient sample is highly dependent on postmortem damage. While in living organisms, DNA damage is repaired by various enzymatic mechanisms, the DNA molecules begin a progressive decay once the metabolic pathways of a cell cease to operate. The decay rate is influenced by a variety of factors related to the environment and the storage conditions. Biochemical processes subsequent to cell death cause the reduction of nucleotide sequence information in many ways which include breakage of the DNA into small fragments, fragmentation of bases and sugars, loss of amino groups and so on (Pääbo *et al.*, 2004). The most common of these modification is the hydrolytic loss of amino groups from the bases adenine, cytosine, 5-methylcytosine and guanine, resulting in hypoxanthine, uracil, thymine and xanthine respectively. The deamination products of cytosine (uracil), 5-methyl-cytosine (thymine) and adenine (hypoxanthine) are of particular relevance for the amplification of ancient DNA since they cause incorrect bases to be inserted when new DNA strands are synthesized by a DNA polymerase. These kinds of PCR artifacts, termed as miscoding lesions are commonly represented by 2 types of transitions: $(A \rightarrow G)/(T \rightarrow C)$ and $(C \rightarrow T)/(G \rightarrow A)$ (Hansen *et al.*, 2001).

With the improvement in amplification techniques, number of such artifacts has reduced, but the precise rate or pattern of occurrence of miscoding lesions are still unknown. Hofreiter *et al.* (2001) calculated the approximate rate of postmortem damage by comparing the PCR products of ancient samples with a database of reference sequences. He concluded that miscoding lesions are unlikely to be more frequent than 0.1%. A study performed by Briggs *et al.* (2007) on ancient DNA samples to investigate the patterns of nucleotide mis-incorporations shows that substitutions resulting from miscoding cytosine residues are vastly overrepresented in the DNA sequences and drastically clustered in the ends of the molecules, whereas other substitutions are rare. According to Gilbert *et al.* (2007), the inflated rate of transitions attributed to DNA damage processes could be due to inclusion of actual PCR errors to DNA damage.

Depurination which is the loss of a purine base (A or G) usually due to an unstable bond between purine bases and the backbone sugar is also one of the principal forms of damage to ancient DNA in fossil or sub-fossil material. Depurinated bases in double-stranded DNA are efficiently repaired by portions of the base excision repair pathway but depurinated bases in single-stranded DNA undergoing replication can lead to mutations. This is because, in the absence of information from the complementary strand, an incorrect base can be added at the apurinic site. According to Briggs *et al.* (2007) depurination causes overrepresentation of purines at positions adjacent to the breaks in the ancient DNA.

1.8.2 PCR Errors

The use of PCR to amplify a DNA target and clone has become an important process in molecular biology. Applications of PCR are diverse and in certain cases its use is critical for example in the field of forensic science, studies on ancient DNA or for estimating microbial diversity, where either a very small amount of DNA is available or DNA is degraded. PCR has been successfully used in all these fields however the validity of results depend highly on PCR fidelity. Due to inherent problem, PCR may produce sequence copies which contain errors. The rate of PCR errors is not negligible, according to Kobayashi *et al.* (1999), approximately 10% of all sequences contain one or more PCR errors when a typical 250 bp sequence is amplified. Most of the work on the PCR errors has been done in the context of microbial biodiversity. Small-Subunit (SSU) *rRNA* genes represent native microbial species and PCR has become a popular tool for retrieval from natural environments of (SSU) *rRNA* genes. The appearance of PCR artifacts is a potential risk in the PCR-mediated analysis of complex microbiota as it suggests the existence of organisms that do not actually exist in the sample investigated (Wintzingerode *et al.*, 1997). There are two main types of errors associated with PCR (Acinas *et al.*, 2005): PCR induced point mutations and formation of chimerical molecules.

PCR-generated mutations are a potential problem for accurate determination of sequence diversity. The major cause of such mutations is Taq DNA polymerase which has a higher intrinsic misincorporation rate during synthesis (Cline *et al.*, 1996). Such errors can accumulate and be enlarged during PCR amplification. Most commercially available Taq polymerases is reported to introduce errors at the rate of approximately 10^{-5} to 10^{-6} point mutations/bp/duplication but PCR amplification with the proofreading DNA polymerase from the hyperthermophilic archaeon Pyrococcus furiosus (Pfu) leads to a 10 times improvement in the misincorporation rate as compared to Taq DNA polymerase, which lacks the proofreading activity (Clarke *et al.*, 2001).

The presence of misincorporated nucleotides is highly problematic when they are located at sites which have been selected as a probe target or when small differences in sequence are used for discrimination. The major problem with PCR induced mutations is that if a mutation occurs during the early cycles of PCR, it is replicated hundreds and thousands of times and finally we may see some closely related amplicons with difference of only 1 or 2 nucleotides. At this point, it becomes difficult to say that if some of these closely related sequences are generated due to PCR error or all of the sequences are genuine. Cummings *et al.* (2010) proposed a method based on binomial distribution for calculating the probability of detecting a given number of PCR artifacts in an amplicon, and thus identify sequences with likely base misincorporations. This method calculates the probability using the following formulae:

$$P(x \geq k) = 1 - \sum_{i=1}^{k} \binom{N}{k-i} \mathcal{E}^{k-i} (1-\mathcal{E})^{N-(k-i)}$$

Here k is the number of PCR errors, N is the total number of bases in the sequence and $\mathcal{E}$ is the PCR error rate in the entire amplicon. The probability is compared with Bonferroni corrected critical value in order to measure the likelihood of an amplicon being a PCR artifact. However according to the author this method is appropriate only for studies involving genes with low genetic diversity.

Chimeras are clones that contain adjacent DNA stretches which are normally located at two very different sites within a genome that is to be sequenced. Chimeras between two different DNA molecules with high sequence similarity (*i.e.* homologous genes) can be generated during PCR process, as DNA strands compete with specific primers during the annealing step. If chimeras are not recognized, this can lead to wrong interpretation of the sequenced organisms. Several studies have been done to estimate the chimeric formation and suggestions have been given to avoid them. According to (Wang and Wang, 1996) chimera formation can be decreased with increasing elongation time, when mixtures of two different 16S rRNA genes were amplified concluding that frequency of chimeric products is positively correlated with number of PCR cycles and sequence similarity between mixed templates. It has also been observed that in addition to incomplete strand synthesis during the PCR process, DNA damage promotes the formation of chimeric molecules. According to Pääbo *et al.* (1990) all kinds of DNA damage including template breaks, UV irradiation and depurination support production of recombinant PCR products.

Several methods have been developed for detecting chimeric sequences which use different methodologies. For example, the Ribosomal Database Project (RDP-II) developed by Cole *et al.* (2003) provides a program called Chimera Check and Komatsoulis and Waterman (1997) has developed an application called chimeric alignment to detect chimeric sequences. Both of these programs rely on direct comparison of individual sequences to one or two putative parent sequences at a time. Other existing algorithms include Pintail by Ashelford *et al.* (2005) and Bellerophon developed by Huber *et al.* (2004). These two programs were developed for removing chimeras from full length clone sequences and lack the sensitivity for short sequence reads. More recent applications developed to detect chimeric reads include; ChimeraSlayer (Haas *et al.*, 2011) and Persus (Quince *et al.*, 2011). These two applications are developed to detect chimeras from short pyrosequencing reads. ChimeraSlayer requires a reference data set of sequences that are known to be non-chimeric, however Persus treats the problem of chimeric detection as a ′classification′ or ′supervised learning′ problem and thus does not require a set of reference sequences.

1.8.3 Sequencing Errors

A sequencing error also termed as "mis call" occurs when a sequencing method calls one or more bases incorrectly leading to an inaccurate read. No sequencing method is perfect and all of the available techniques produce errors occasionally. However, the chance of a sequencing error is generally known and quantifiable. This is done by assigning a quality score to each base in the read, indicating confidence that the base has been called correctly. Some sequencing methods are more reliable than others and so give higher quality scores. Generally sequencing errors are more likely to appear at the end of a read. Sequencing errors can be traced by aligning the sequenced read with the reference sequence (reference sequence is considered as the actual true sequence) and observing the differences. Two different types of sequencing errors are normally observed.

- Mismatches: A mismatch is a substitution of one base for another, *e.g.* an A for a C.
- Indels: The word indel is an abbreviated form for "insertion/deletion". This type of errors occur when a read contains a different number of bases from its reference sequence at some points in the alignment. An insertion occurs when the read contains extra bases, while a deletion occurs when the read is missing a base.

Traditional sequencing was based on Sanger′s method (Sanger *et al.*, 1977). This method can sequence up to 1000 bp long reads at an error rate as low as 10^{-5} error per base (Shendure and Ji, 2008). The continuous demand for cheap and fast sequencing technology has led to the development of next generation sequencing technologies which improve the sequencing speed and lower the cost at the price of a lower accuracy and shorter read lengths compared to Sanger sequencing. While the overall production pipelines are similar across different sequencing platforms, they differ in mechanistic details which affect the types of errors made during sequencing. One important step during the sequencing process is of base calling that involves the analysis of sensor data to predict the individual bases. The type of errors produced depend on the base calling procedure as well. The characterization of errors associated with the different sequencing platforms is of crucial importance for sequences analyses. So in order to explain the different types of errors produced by these sequencers, it is important to understand their sequencing chemistry. In this regard we will consider the two latest sequencing technologies mentioned above *i.e.* 454 pyrosequencing and Solexa system. We will see how the sequencing process differs and how the base calling is performed for these two techniques with respect to the types of errors produced by them. A detailed comparison of all of next generation sequencing techniques with respect to the type errors generated by them is given in (Shendure and Ji, 2008).

454 Pyrosequencing

The 454 pyrosequencing process uses a sequencing by synthesis approach to generate sequence data. Sequencing by synthesis approach involves serial extension of primed templates. A long double helix DNA molecule is broken down into shorter fragments of approximately 400 to 600 base pairs and adapter molecules are attached to short DNA fragments. The adapter molecules help in amplification and sequencing process. Next the adapter flanked double stranded DNA fragments are separated into single strands and fixed on small DNA-capture beads. The DNA fixed to these beads is amplified by emulsion PCR in order to increase the downstream signal intensity. Ideally, during this process a single template is attached to each bead leading to uniform clusters on each bead. During the PCR, a single DNA fragment is amplified into approximately ten million identical copies that are immobilized on the capture beads. When the PCR reaction is complete, the beads are filtered eliminating the beads which do not hold any DNA. The beads are then deposited onto an array of picoliter-scale wells (Margulies *et al.*, 2005) such that each well contains a single bead. At this point some enzymes like polymerase and luciferase are also added which help in the synthesis and detection process. Finally the PicoTiterPlate is placed into the 454 GS FLX/Roche System for sequencing. The sequencing process consists of alternating cycles of enzymes driven biochemistry and image processing of data produced (Shendure and Ji, 2008). The 454 instrument includes a fluidics system capable of washing the PicoTiterPlate with various reagents including the A, C, G and T nucleotides. The four nucleotides are flowed sequentially over the PicoTiterPlate (the process is repeated almost 100 times for a large run). When a complementary nucleotide enters a well, the template strand is extended by DNA polymerase. At this point the bead-bound enzymes contained in each PicoTiterPlate well convert the chemicals generated during nucleotide incorporation into light in a chemi-luminescent reaction. This light is detected by CCD sensors in the instrument. The intensity of light generated during the flow of a single nucleotide is proportional to the consecutive number of complementary nucleotides incorporated on the single stranded DNA fragment. For example, if there are three consecutive T's in the single-stranded fragment, the amount of light generated would be three times that of a single T in the fragment. The result of this sequencing process is a flowgram showing the intensities of light produced at each incorporation of a nucleotide. Base calling is done by reading the flowgram and putting threshold values to determine the number of consecutive bases at a point. A sample of flowgram is shown in Figure 1.5.

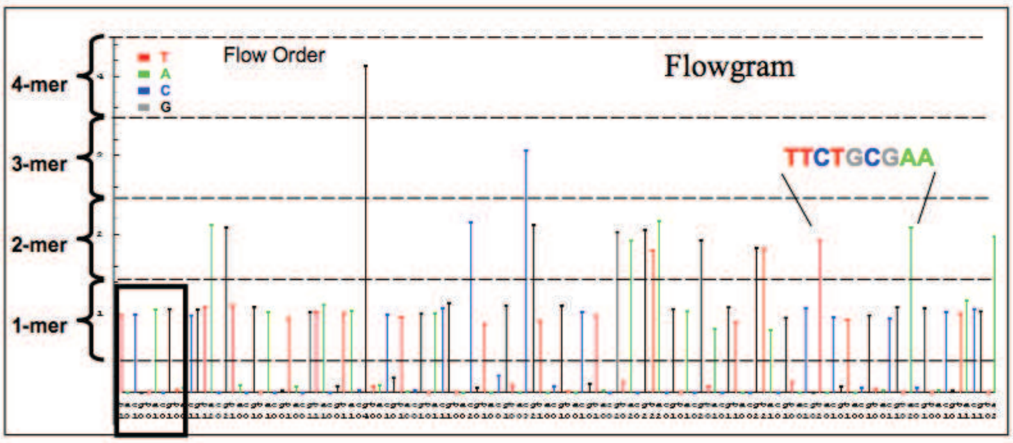


Figure 1.5: Bar graph of light intensities called a flow-gram for each well contained on the PicoTiterPlateTM. The signal strength is proportional to the number of nucleotide incorporated.

A major limitation of 454 technology is related to homopolymers that is, consecutive instances of the same base, such as CCC or AAAA. Since all bases of a homopolymer are included in a single cycle, the length of a homopolymer is inferred from the signal intensity which is prone to a greater error rate. The standard base-calling procedure rounds off the continuous intensities to integers. Consequently, long homopolymers result in frequent miscalls: either insertions or deletions, which is the dominant error type for 454 technology.

During the base calling a quality score is assigned to every called base. More commonly used quality scores are Phred scores (Ewing and Green, 1998) which define the quality value q assigned to a base-call to be:

$$q = -10 \times \log_{10}(p) \tag{1.8.1}$$

where p is the estimated error probability for that base-call. So a base-call having a probability of 1/1000 of being incorrect is assigned a quality value of 30. This means high quality values correspond to low error probabilities, and vice versa. This quality score corresponds to the log probability that the base was not an overcall, that is, the predicted homopolymer length was not too long.

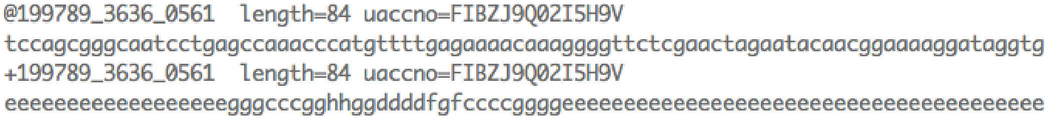
Illumina/solexa

This platform was first introduced by Solexa in 2006 and later on re-branded as Illumina Genome Analyzer (GA). GA is a sequencing by synthesis technology and supports massively parallel sequencing using a reversible terminator-based method which enables detection of single bases as they are incorporated into growing DNA strands.

The Genome Analyzer uses a flow cell consisting of an optically transparent slide with 8 individual lanes such that eight independent samples/libraries can be sequenced in parallel during the same instrument run. Single stranded oligonucleotide anchors are bound on the surface of flow cell. Libraries can be prepared by any method that gives rise to a mixture of adaptor-flanked fragments of size ranging from $150 - 200$ bp. These adapter-flanked oligonucleotides are complementary to the flow-cell anchors. Adapter-flanked template DNA is added to the flow cell and immobilized by hybridization to the anchors. DNA templates are amplified in the flow cell by bridge amplification, which relies on captured DNA strands arching over and hybridizing to an adjacent anchor oligonucleotide. Multiple amplification cycles convert the single-molecule DNA template to a clonally amplified arching cluster with each cluster containing approximately 1000 clonal molecules. Almost 60×10^6 separate clusters can be generated per flow cell. After cluster generation, the amplicons are single stranded and sequencing is initiated. Sequencing is done by hybridizing a primer complementary to the adapter sequences followed by addition of polymerase and a mixture of 4 differently colored fluorescent nucleotides. These nucleotides are reversible terminators which means that a chemically cleavable moiety at the 3′ hydroxyl position allows only a single-base to be incorporated in each cycle. Fluorescent emission identifies which of the four bases was incorporated at that position. After a single-base extension fluorescent emission is recorded by taking image. With some chemical steps, the reversible terminator nucleotides are unblocked, the fluorescent labels are cleaved and washed away, and the next sequencing cycle is performed.

The typical read length is 100 bp but read length is inversely related to base calling accuracy (Dohm *et al.*, 2008). Read lengths are limited by multiple factors such as incomplete cleavage of fluorescent labels or terminating moieties and under- or over-incorporation of nucleotides. With successive cycles these errors can be accumulated producing a heterogeneous population. The dominant error types in Solxa are substitutions rather than insertions or deletions and homopolymers are less of an issue with this technique (Shendure and Ji, 2008). Illumina platform also provides a quality score with each base-call like those of Phred quality scores. Although average raw error rates are on the order of $1 - 1.5\%$, but higher accuracy bases with error rates of 0.1% or less can be identified through quality scores provided with each base-call. Recently Illumina has also started using "Paired-end" strategy to sequence both ends of template molecules. This strategy provides positional information that facilitates alignment especially for short reads (Korbel *et al.*, 2007).

Sequence File Formats Sequence reads are provided in special formats from vendors, the most commonly used is FASTQ format which has recently become the de facto standard for storing the output of high throughput sequencing instruments. FASTQ files have both sequence and its corresponding quality value (Phred quality score). This format normally uses four lines per sequence. Line 1 begins with a '@' character and is followed by a sequence identifier and an optional description (like a FASTA title line). Line 2 is the raw sequence letters. Line 3 begins with a '+' character and is optionally followed by the same sequence identifier and any description again. Line 4 encodes the quality values for the sequence in Line 2 and must contain the same number of symbols as letters in the sequence. Quality scores are encoded with ASCII characters, however encoding schemes varies from vendor to vendor. Sanger format encodes a Phred quality score from 0 to 93 using ASCII 33 to 126 and Illumina format (version 1.3+) encodes a Phred quality score from 0 to 62 using ASCII 64 to 126. A sample FASTQ format is shown in Figure 1.6.



```
@199789_3636_0561  length=84 uaccno=FIBZJ9Q02I5H9V
tccagcgggcaatcctgagccaaacccatgttttgagaaaacaaaggggttctcgaactagaatacaacggaaaaggataggtg
+199789_3636_0561  length=84 uaccno=FIBZJ9Q02I5H9V
eeeeeeeeeeeeeeeeeegggcccgghhggddddfgfccccggggeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeee
```

Figure 1.6: An example of FASTQ format for representing sequences and their quality scores in ASCII encoding

Dealing With Sequencing Errors

Since different types of errors are produced by the two latest sequencing technologies, thus it is required to have different error models each specific to the particular sequencing type. However most of the work in order to deal with sequencing errors has been done in the context of accurate determination of microbial diversity from 454 pyrosequencing. The most well-known among these algorithms being perhaps PyroNoise (Quince *et al.*, 2009) that differentiates noisy reads from true sequences of a 454 run. PyroNoise reconstructs true sequences and frequencies in the sample prior to OTU construction. It is based on clustering flowgrams rather than sequences which allows 454 errors to be modeled naturally. Pyronoise uses light intensities associated with each read or flowgram and defines a distance reflecting the probability that a flowgram was generated by a given sequence. The distances are used in a mixture model to define a likelihood of observing all the flowgrams assuming that they were generated from a set of true underlying sequences. The program uses an iterative expectation-maximization algorithm to maximize this likelihood and obtain the de-noised sequences. The algorithm first calculates the most likely set of sequences given the probabilities that each flowgram was generated by each sequence and then recalculates those probabilities given the new sequences. The procedure is then repeated until the algorithm converges. Another similar algorithm based on flowgram de-noising is DeNoiser, that has been developed by Reeder and Knight (2010). This algorithm also uses flowgram alignment technique as used by PyroNoise, however it uses a greedy agglomerative clustering approach instead of iterative approach as used by Pyronoise and thus it reduces the computational cost of algorithm. The algorithm starts by finding unique sequences, orders them by frequency and then starting with the most abundant sequence (called centroid), maps the other reads onto these centroids if their distance to the centroid is smaller than some threshold. However with this approach mis-assignments of reads is quite probable when true sequences are very similar and thus it is possible that OTUs are not accurately reconstructed. Another similar approach is the centroid based clustering using sequence distances rather than flowgrams based distances, called single linkage pre-clustering (Huse *et al.*, 2010). This strategy has been used in the program PyroTagger developed by Kunin and Hugenholtz (2010). Recently a new program AmpliconNoise (Quince *et al.*, 2011) has been developed that is capable of separately removing 454 sequencing errors and PCR single base errors and can detect chimera too. AmpliconNoise is an extension of Pyronoise algorithm, in this new program flowgram clustering is performed without alignment followed by an alignment based sequence clustering. Sequence clustering accounts for the differential rates of nucleotide errors in the PCR process, and uses sequence frequencies to inform the clustering process. The result of this approach is a lower computational cost because the fast alignment free flowgram clustering reduces the data set size for the slower sequence clustering.

Some programs are also available for short read data analysis including reads output from both 454 and Solxa system. ShortRead package (Morgan *et al.*, 2009), part of the Bioconductor project is one such program that provides tools for quality assessment and data transformation *etc*. However according to our knowledge no program is available that can model errors generated by Illumina/solexa technology.

1.9 Conclusion

In this chapter we have covered most of the important topics related to the concept of species inventory and DNA barcoding in the context of biodiversity assessment. There are three main barcoding challenges faced by scientific community today: evaluating the quality of barcode regions to chose the better markers, designing new optimal barcode markers and their corresponding primers and the analysis of huge amount of data produced by next generation sequencing projects to understand errors behavior. Each of this task is of utmost importance particularly for ecological studies related to ancient DNA. We have presented all important background information related to these three tasks in this chapter so that the following chapters are easier to comprehend.

1.10 Résumé

Durant les dix dernières années, le technique dite du code barre ADN s'est imposée comme une méthode de choix pour l'identification rapide de spécimens biologique. Nous pouvons essayer d'allez un peu plus loin en définissant deux principaux types de "DNA barcoding" : le DNA barcoding conventionnel telqu'il est défini par le "Consortium for the Barcoding of Life" et qui vise à identifié un spécimen biologique avec une précision taxonomique à l'espèce et le DNA barcoding *sensu lato* ou DNA metabarcoding utilisé pour l'identification simultanée de tous les organismes présent dans un écosystème.

Les écologistes ont souvent besoin de déterminer la liste des espèces impliquées dans le processus écologique qu'ils étudient. L'établissement de cette liste est le généralement une tache ardue requérant un effort d'échantillonnage important et une forte compétence taxonomique. En couplant le principe du barre code ADN aux dernières technologies de sequençage haut débit, le DNA metabarcoding peut produire une grande quantité de données permettant de mesurer la biodiversité. Mais malgré sa relation théorique avec la technique du code barre ADN, les specificités du DNA metabarcoding font qu'il est nécessaire de développer des outils spécifiques tant pour le choix des marqueurs que pour l'analyse des données. Dans ce contexte, cette thèse a été consacrée au développement de techniques bioinformatiques facilitant l'utilisation du DNA metabarcoding pour une évaluation précise de la biodiversité.

Le premier chapitre d'introduction et d'état de l'art de cette thèse couvre deux grands domaines: le première, plus biologique, aborde les sujets liés à l'importance des systèmes de classification biologique utilisé pour l'inventaire d'espèces et leur l'identification, les méthodes de classification bien connu et une introduction détaillée au DNA barcoding et au DNA metabarcoding. La deuxième partie plus technique introduit certains termes d'informatique et donne un aperçu des principaux algorithmes de recherche de répétition dans les chaînes de caractères avec quelques détails sur leur complexité.

CHAPTER 2

Formal Measures For Barcode Quality Evaluation

2.1 Introduction

From the discussion of first chapter, it is clear that different metabarcoding applications may require different barcode markers. Since there exist no standard markers for metabarcoding, we need to design them. However, an important question in this regard is how to estimate the quality of these markers and how to chose a suitable marker from a set of markers proposed for metabarcoding applications? This question leads towards a kind of inference problem, where, the inference algorithm should be able to compare two solutions to decide which one is the best. The simplest strategy to many inference problems is to define a score function where possible, in order to choose the best solution. The scoring function is also called an objective function and this function can be defined by making use of knowledge about a specific subject and introducing this knowledge into the algorithm. A score function depends on some parameters and observations.

In order to define a score function for measuring the quality of barcode regions, we can introduce our knowledge about good barcode markers into our algorithm. We know that an ideal barcode maker should be able to amplify as many taxa as possible and it should be able to well discriminate among different taxa. Thus the quality of a barcode region mainly depends on two factors.

- The ability of the primers to amplify a broad range of taxa.
- The ability of the region to discriminate between two taxa.

We can also define a third measure *i.e.* length of the barcode marker. This length constraint is important, because, in case of environmental applications where DNA is degraded, the smallest possible amplifiable regions are preferred. Depending on the application, we may be interested in optimizing one or all of the quality measures *i.e.* amplification range, taxa discrimination capability and the length of a barcode region.

Based on the two factors mentioned above, we have developed two formal measures *i.e.* barcode $coverage$ (B_c) and barcode $specificty$ (B_s) to score a barcode region. B_c gives a quantitative measure of amplification range and B_s gives a quantitative measure of taxa discrimination capacity of a barcode region. These two measures were published by Ficetola *et al.* (2010). This article is dedicated to comparison of several metabarcode markers for vertebrates. I participated to this work by developing the two indices described above.

2.2 An *In silico* Approach For The Evaluation Of DNA Barcodes

We used these two indices to measure the relative quality of standard barcode markers in the context of metabarcoding applications. In this article we also present our program $ecoPCR$ that performs *in silico* PCR for a selected primer pair on a large sequence database. Using the definition of proposed quality indices and processing the output of $ecoPCR$ program, we compared the taxonomic coverage and resolution of several DNA regions already proposed for the barcoding of vertebrates. The publication follows on the next page.

BMC Genomics

METHODOLOGY ARTICLE Open Access

An *In silico* approach for the evaluation of DNA barcodes

Gentile Francesco Ficetola[1,2,3*†], Eric Coissac[1*†], Stéphanie Zundel[1], Tiayyba Riaz[1], Wasim Shehzad[1], Julien Bessière[1], Pierre Taberlet[1], François Pompanon[1]

Abstract

Background: DNA barcoding is a key tool for assessing biodiversity in both taxonomic and environmental studies. Essential features of barcodes include their applicability to a wide spectrum of taxa and their ability to identify even closely related species. Several DNA regions have been proposed as barcodes and the region selected strongly influences the output of a study. However, formal comparisons between barcodes remained limited until now. Here we present a standard method for evaluating barcode quality, based on the use of a new bioinformatic tool that performs *in silico* PCR over large databases. We illustrate this approach by comparing the taxonomic coverage and the resolution of several DNA regions already proposed for the barcoding of vertebrates. To assess the relationship between *in silico* and *in vitro* PCR, we also developed specific primers amplifying different species of Felidae, and we tested them using both kinds of PCR

Results: Tests on specific primers confirmed the correspondence between *in silico* and *in vitro* PCR. Nevertheless, results of *in silico* and *in vitro* PCRs can be somehow different, also because tuning PCR conditions can increase the performance of primers with limited taxonomic coverage. The *in silico* evaluation of DNA barcodes showed a strong variation of taxonomic coverage (i.e., universality): barcodes based on highly degenerated primers and those corresponding to the conserved region of the *Cyt*-b showed the highest coverage. As expected, longer barcodes had a better resolution than shorter ones, which are however more convenient for ecological studies analysing environmental samples.

Conclusions: *In silico* PCR could be used to improve the performance of a study, by allowing the preliminary comparison of several DNA regions in order to identify the most appropriate barcode depending on the study aims.

Background

DNA barcoding, i.e., the identification of biological diversity using standardized DNA regions, has been demonstrated as a new, very useful approach to identify species [1]. Originally, DNA barcoding was proposed to assign an unambiguous tag to each species, giving to taxonomists a standard method for identification of specimens. In this context, it was also proposed that DNA barcoding is an opportunity to accelerate the discovery of new species [2-4]. Today, the fields of applications of this approach are broader. As example, DNA barcoding has been already used in biodiversity assessment, forensics, diet analysis and paleoecological studies [5-7].

In the former context, a portion of mitochondrial cytochrome *c* oxidase (*COI*) has been proposed as the standard barcode for animal identification [1,8]. Since then, other portions of DNA have been proposed as barcodes, because different DNA regions have different performances in some taxa (e.g., flowering plants [9,10]; amphibians [11]). If we consider the other applications of barcoding (*sensu lato* DNA barcoding, [6]), the necessity to limit the number of usable barcode loci for conserving the standard aspect of this method can be relaxed. In such a new context, multiple barcodes in different regions of the genome could be combined to improve identification, according to the taxon studied and to the aims of the research [9,10]. Therefore, the first step of a *sensu lato* barcoding study should be the selection of the best DNA region(s) to be used as barcode considering the aims of the study. The availability of large public sequence databases may allow comparing multiple potential barcodes and their properties before performing studies.

Among the properties of an ideal DNA barcode, high taxonomic coverage and high resolution are essential [6,12]. A high taxonomic coverage (also called universality) would allow the application of barcodes to a number of taxa as large as possible, including undescribed species. This constraints the DNA barcode region to have sufficiently conserved flanking regions enabling the design of universal primers. This is especially important for describing unknown biodiversity or diversity within environmental samples such as soils or faeces [6,7,13]. However, universality can be extremely difficult to achieve, because of the incomplete knowledge of genetic variation in poorly studied taxa [12]. The resolution capacity of a barcode is its ability to differentiate and identify species that relies on interspecific differences among DNA sequences [8,14]. Thus, the challenge for defining a barcode of good quality consists in finding a quite short and enough variable DNA sequence flanked by highly conserved regions. Depending of the application, the size, the taxonomic coverage or the resolution of the DNA barcode could be the most important characteristic to optimise [6].

This study proposes an explicit approach for comparing the performance of potential barcoding regions, which is based on '*in silico* PCRs' performed over extensive databases, and on two indices that estimate the resolution capacity of the barcodes and the taxonomic coverage of the primers used for their amplification. As an example, we analysed several primers available from the literature that have been used in *sensu lato* barcoding studies [6] for the identification of Vertebrates species. First, we assessed the taxonomic coverage of several primer pairs by evaluating the proportion of species amplified *in silico* in a purposely designed database. Subsequently, we analyzed the GenBank sequences amplified by each primer pair, in order to evaluate the proportion of species correctly identified on the basis of their barcodes. We also used an *in vitro* analysis to validate the correspondence between *in silico* and real world PCR.

Methods

General strategy

First, we created a reference database representative of the mitochondrial genomes of all vertebrates, by retrieving from Genbank all the complete mitochondrial genomes of Vertebrates available (accession: September 2007). Subsequently, we randomly selected one sequence per species, to reduce the overrepresentation of a few species (e.g., humans, mouse, zebrafish etc.). We obtained a set of 814 mitochondrial genomes representative of the five major monophyletic clades of vertebrates [Chondrichthyes: 8 species; Actinopterigii: 385 species; Amphibia: 79 species; Sauropsida (= birds + "reptiles"): 133 species; Mammalia: 202 species; other taxa: 7 species]. Most of species were the unique representative of their genus and the database corresponded to 633 genera.

To analyze the performance of each primer pair studied, we first performed an *in silico* PCR on the reference database and we evaluated the taxonomic coverage of each primer pair as the proportion of amplified taxa. Then, we performed an *in silico* PCR on the whole GenBank, to evaluate the resolution of the amplified fragments that represents the proportion of unambiguously identified taxa. These properties were evaluated for the whole Vertebrates and for each of the five clades which compose it.

In Silico PCR

An *in silico* PCR consists in selecting in a database the sequences that match (i.e., exhibit similarity with) two PCR primers. The regions matching the two primers should be localised on the selected sequence in a way allowing PCR amplification, which forces the relative orientation of the matches and the distance between them. In order to simulate real PCR conditions, the *in silico* PCR algorithm should allow some mismatches between the primers and the target sequences. Standard sequence similarity assessment programs such as BLAST [15] are not suitable for such kind of analysis because the heuristic search they use is not efficient on short sequences. Moreover, a post processing of BLAST output should be performed to verify previously stated constraints. We have developed a program named ecoPCR that is based on the very efficient pattern matching algorithm Agrep [16]. This algorithm allows specifying the maximum count of mismatched positions between each primer and the target sequence, and to use the full IUPAC code (e.g., R for purines or Y for pyrimidines). It also allows specifying on which primer's specific positions mismatches are not tolerated, what is useful to force exact match on the 3′ end of primers for simulating real PCR conditions. Moreover, to facilitate further analysis, ecoPCR output contains the taxonomic information for each sequence selected from the database. For the analyses presented in this article, we allowed two mismatches between each primer and the template, except on the last 3 bases of the 3′ end of the primer. Analyses performed with 0, 1 or 3 mismatches led to similar conclusions (results not shown), even if the results were sometimes different (see discussion). This software was developed for Unix platforms and is freely available at http://www.grenoble.prabi.fr/trac/ecoPCR.

Measuring taxonomic coverage

To measure the taxonomic coverage of a primer pair, we defined a coverage index B_c as the ratio between the number of amplified taxa for a specified taxonomic rank (i.e., species for this analysis; genus or family can be specified as alternative taxonomic ranks) and the total number of taxa of the same level representing the studied clade in the reference sequence database. B_c can be computed from ecoPCR output file using the ecoTaxStat script.

Measuring resolution capacity

The resolution capacity of a barcode was estimated by an index measuring the ratio of unambiguously identified taxa for a given taxonomic level over the total number of tested taxa. A taxon unambiguously identified by a primer pair owns a barcode sequence associated to this pair that is not shared by any other taxa of the same taxonomic rank. To be computed, this definition can be formalized considering the mapping *E*, *Img* and *E′* between four concept sets: taxon (*T*), individual (*I*), barcode (*B*) and region (*R*) (for a full definition see figure 1). Considering the a taxon $t \in T$ and a primer pair (barcode region) $r \in R$ and using the mapping *E*, *Img* and *E′* we define the Ω(t,r) set of all barcodes belonging to a taxon for a region:

$$\Omega(\mathrm{t},\mathrm{r}) \ = Img(E(t)) \ \cap E'(r)$$

From the above description, we note the set of all individuals owning a barcode corresponding to a taxon as:

$$Img^{-1}(\Omega) \equiv \bigcup_{i} Img^{-1}(b_i \,/\, b_i \in \Omega)$$

This allows defining an unambiguously identified taxon t by a barcode region r if and only if:

$$Img^{-1}(\Omega(t,r)) = E(t)$$

This defines a mapping *ε* of *T* to *R* and allows to define the specificity index B_s as:

$$B_s(r) = \frac{|\{t/t\varepsilon\ r\}|}{|T|}$$

B_s can be computed from an ecoPCR output file using the ecoTaxSpecificity script. ecoTaxSpecificity and ecoTaxStat scripts are parts of the OBITools python package freely available at http://www.grenoble.prabi.fr/trac/OBITools.

In a few cases, especially for Chondrichthyes, ecoPCR ran over the entire GenBank yielded only a small number of sequences. Thus, we calculated the resolution capacity of a barcode only when the primer pair amplified more than 10 species.

Correspondance between in vitro and in silico PCRs

Strict experimental validation of the electronic PCR realized over large databases would be extremely difficult, as it would require obtaining tissues from hundreds of species. Alternatively, specific primer pairs designed to amplify only one species can be used to confirm the correspondence between the results of ecoPCR and *in vitro* PCR. Therefore, we designed specific primers to amplify mitochondrial DNA of three species, using ecoPCR to test their specificity. Then, we cross-amplified the three species with each primer pairs with *in vitro* PCR to verify the ecoPCR predictions.

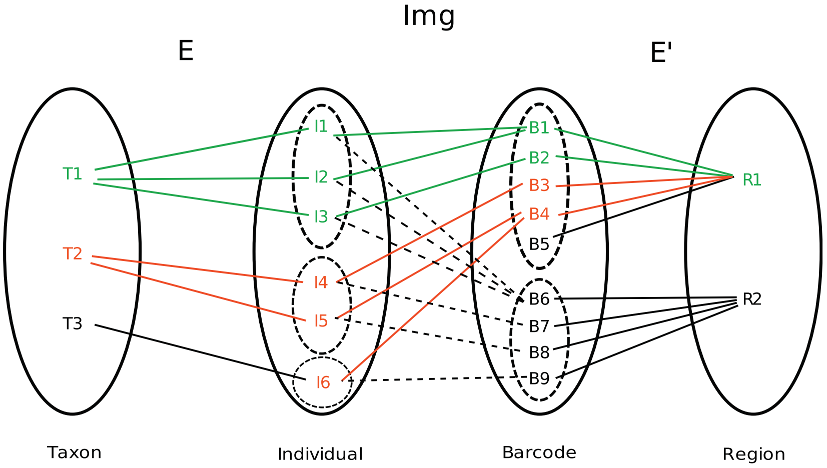


Figure 1 Relationships between taxa, individuals, barcodes and regions as used in the B_s index estimation. In this example the taxon T1 is unambiguously identified by the R1 barcode region (green links) but the T2 is not well identified by the R1 region because this taxon share the B4 barcode region with the T3 taxon via the I6 individual (red links).

We considered three species of Asiatic Felidae: the Leopard (*Panthera pardus*); the Snow Leopard (*Uncia uncia*) and the Leopard cat (*Prionailurus bengalensis*). We designed specific primers for amplifying short sequences of mitochondrial 12S; this kind of primer pairs can be used to identify species from degraded DNA and remains, such as faeces. The three primer pairs were: (a) *Pant*F, 5′-GTCATACGATTAACCCGG-3′; *Pant*R, 5′-TGCCATATTTTTATATTAACTGC-3′, designed to amplify the Leopard (amplified fragment: 120 bp); (b), *Unci*F, 5′-CTAAACCTAGATAGTTAGCT-3′, *Unci*R, 5′-CTCCTCTAGAGGGGTG-3′, designed to amplify the Snow Leopard (amplified fragment: 104 bp); (c) *Prio*F, 5′-CCTAAACTTAGATAGTTAATTTT-3′, *Prio*R, 5′-GGATGTAAAGCACCGCC-3′, designed to amplify the Cat Leopard (amplified fragment: 94 bp). DNA was extracted from faeces using QiAamp DNA Stool Kit (Qiagen GmbH, Hilden, Germany). The PCRs were conducted in a 20 μl total volume with 8 mM Tris-HCl (pH 8.3), 40 mM KCl, 2 mM MgCl2, 0.2 μM of each primer, BSA (5 μg), 0.5 U of AmpliTaq Gold DNA polymerase (Applied Biosystems) and 2 ml of DNA extract. For all primers, the PCR programme included an initial 10 min denaturation step at 95°C, 45 cycles of denaturation at 95°C for 30 s and annealing at 53°C for 30 s. Samples of each of the three species were amplified with the three primer pairs, to verify *in vitro* the possibility of cross-amplification. We also tested cross-amplification ability of these primer pairs using ecoPCR, allowing two mismatches between each primer and the template, except on the last 3 bases of the 3′ end of the primer; subsequently, we simulated more relaxed PCR conditions [17] by allowing a larger number of mismatches.

Vertebrate primer pairs tested

The vertebrate primers tested (table 1) were selected in the bibliography as representative of the diversity of the strategies used for defining barcodes. Some of them (COI-1, COI-2, COI-3) were highly degenerated, in order to maximise the number of taxa amplified (i.e., the taxonomic coverage) [18]. Most of primers chosen amplified long sequences (> 500 bp) to maximize resolution, while some (e.g., Uni-Minibar, 16Smam) have been designed to amplify short sequences, to maximize the possibility of retrieving sequences from damaged/ancient DNA [19-21].

Results

Validation of in silico PCR

With *in vitro* PCR, each pair of specific primers amplified only the species for which it was designed: *Pant*

Table 1 Vertebrate primer pairs tested

Barcode name	Primer Name	Sequence	Fragment size *	Developed for	Reference
COI					
COI-1	FF2d	TTCTCCACCAACCACAARGAYATYGG	655	Fish	[18]
	FR1d	CACCTCAGGGTGTCCGAARAAYCARAA			
COI-2H	LCO1490	GGTCAACAAATCATAAAGATATTGG	658	mainly Arthropods	[1]
	HCO2198	TAAACTTCAGGGTGACCAAAAAATCA			
COI-2	C_VF1LFt1	WYTCAACCAAYCANAANGANATNGG	658	Fish	[18]; modified from [1]
	C_VR1LRt1	TARACTTCTGGRTGNCCNAANAANCA			
COI-3	C_FishF1t1	TCRACYAAYCAYAAAGAYATYGGCAC	652	Fish	[18]
	C_FishR1t1	ACYTCAGGGTGWCCGAARAAYCARAA			
Uni-Minibar	UniMinibarR1	GAAAATCATAATGAAGGCATGAGC	130	Eukaryota	[20]
	UniMinibarF1	TCCACTAATCACAARGATATTGGTAC			
Cyt-*b*					
MCB	mcb398	TACCATGAGGACAAATATCATTCTG	472	All Vertebrates	[30]
	mcb869	CCTCCTAGTTTGTTAGGGATTGATCG			
cytM	L14841	CCATCCAACATCTCAGCATGATGAAA	359	All Vertebrates	[31]; modif. from [26]
	H15149	CCCCTCAGAATGATATTTGTCCTCA			
16S					
16Sr	16Sar	CGCCTGTTTATCAAAAACAT	573	Mammals	[27,28]
	16Sbr	CCGGTCTGAACTCAGATCACGT			
16Sr2	16Sa2	CGCCTGTTTACCAAAAACAT	573	All Vertebrates	this study, modif. from [28]
	16Sb	CCGGTCTGAACTCAGATCACGT			
16Smam	16Smam1	CGGTTGGGGTGACCTCGGA	140	Mammals, ancient DNA	[21]
	16Smam2	GCTGTTATCCCTAGGGTAACT			

* as reported on the original paper.

primers amplified Common Leopard only; *Unci* primers amplified Snow Leopard only, and *Prio* primers amplified Cat Leopard only (Figure 2). Crossamplification through ecoPCR yielded identical results when allowing two mismatches. A more extensive analysis using ecoPCR, and allowing a larger number of mismatches (i.e., simulating more relaxed PCR conditions), shows that *Pant* primers require at least 3 mismatches for cross-amplifying *Uncia uncia*. Similarly, *Unci* and *Prio* primers require at least 4 mismatches for cross amplifying other species.

Evaluation of vertebrate primer pairs: Taxonomic coverage

The primer pairs tested showed very different taxonomic coverage. Overall, COI-2, 16Sr and 16Sr2 were the primers with the highest percentages of species amplified (95, 90 and 93% of vertebrates amplified, respectively; Figure 3, table 2). Following our *in silico* PCRs, the primers with the lowest coverage corresponded to Uni-Minibar, COI-1, COI-2H, MCB and cytM. The primers also differed in their performance in amplifying the major clades of vertebrates. For example, COI-3 had the highest amplification rate in Chondrichthyes, while it amplified only 32% of the mammals. Conversely, 16Smam amplified most of the mammals, but failed in the amplification of Chondrichthyes (Figure 3, table 2). Nevertheless, in a similar way to how modifying the annealing temperature influences *in vitro* PCR [17], the number of electronically amplified species can be quickly increased by allowing a larger number of mismatches (Figure 4). For example, with primers Uni-Minibar, the proportion of amplified species reached 98% with eight tolerated mismatches (Figure 4).

Resolution capacity of barcode regions

When tested over the entire Genbank, most of the primer pairs had a very high resolution capacity, indicated by a high B_s index (Figure 5; table 2). We did not calculate B_s for primers Uni-Minibar and COI-2H because of the low number of species amplified with the settings used for this analysis (see discussion). Only the 16Smam primer pair, which amplifies a very short sequence (140 bp), had $B_s < 85\%$. B_s was ≥ 90% for all other primer pairs and even > 97% for 16Sr and 16Sr2 whatever the vertebrate clade analysed (Figure 3, table 2). Apart from a few cases (e.g., low resolution of cytM within Actinopterigii), the resolution capacity of all primer pairs was consistently high across all taxa tested. These B_s differences are not correlated with the number of Genbank sequences amplified (analysis over all vertebrates: Spearman's correlation $r_S = -0.323$, $N = 8$, $p = 0.4$; the correlations between resolution and number of amplified sequences were not significant also within the monophyletic groups analysed).

The *in silico* PCRs performed over the entire GenBank always yielded sequences from the target mitochondrial region. None of the primers amplified sequences recorded as nuclear sequences in GenBank.

Discussion

The identification of universal primer pairs amplifying fragments with high resolution capacity is a major task

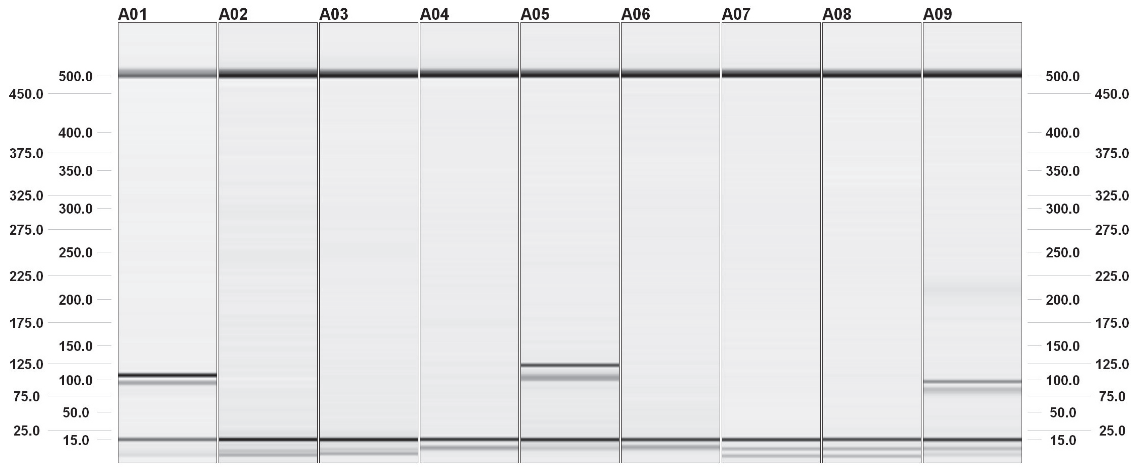


Figure 2 Capillary electrophoresis (QIAxcel System, Qiagen) showing the results of cross amplification of three species of Felidae using three specific primers. A01: *Unci* primers, template DNA from *Uncia uncia*; A02: *Unci* primers, template DNA from *Panthera pardus*; A03: *Unci* primers, template DNA from *Prionailurus bengalensis*; A04: *Pant* primers, template DNA from *U. uncia*; A05: *Pant* primers, template DNA from *P. pardus*; A06: *Pant* primers, template DNA from *P. bengalensis*; A07: *Prio* primers, template DNA from *U. uncia*; A08: *Prio* primers, template DNA from *P. pardus*; A09: *Prio* primers, template DNA from P. bengalensis. The size in base pairs is indicated on the left and on the right.

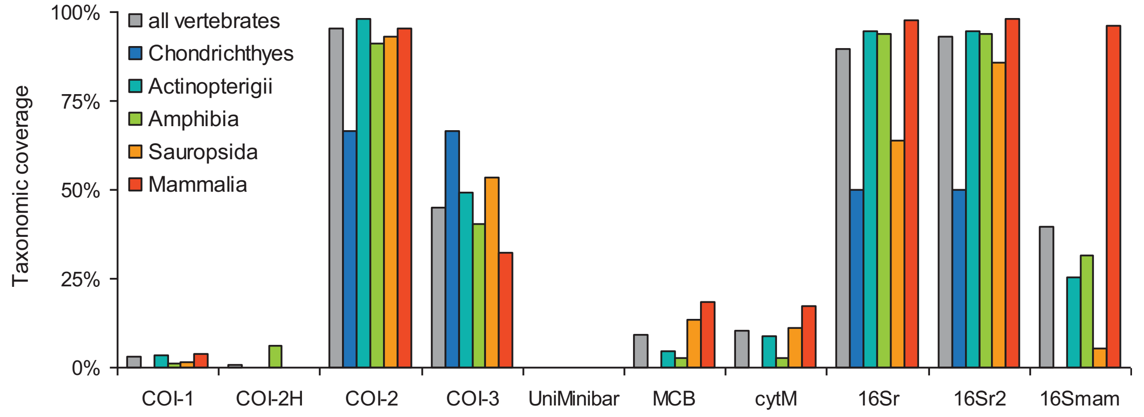


Figure 3 Taxonomic coverage of different primer pairs tested over the reference database.

Table 2 Taxonomic coverage and resolution capacity (B_S) of the different barcodes tested.

	all vertebrates		Chondrichthyes		Actinopterigii		Amphibia		Sauropsida		Mammalia	
					Taxonomic coverage							
COI-1	0.03		0.00		0.03		0.01		0.02		0.04	
COI-2H	0.01		0.00		0.00		0.06		0.00		0.00	
COI-2	0.95		0.67		0.98		0.91		0.93		0.96	
COI-3	0.45		0.67		0.49		0.41		0.53		0.32	
Uni-Minibar	0.00		0.00		0.00		0.00		0.00		0.00	
MCB	0.09		0.00		0.05		0.03		0.14		0.18	
cytM	0.10		0.00		0.09		0.03		0.11		0.17	
16Sr	0.90		0.50		0.94		0.94		0.64		0.98	
16Sr2	0.93		0.50		0.94		0.94		0.86		0.98	
16Smam	0.40		0.00		0.25		0.32		0.05		0.96	
					Resolution capacity							
	B_S	*N*	B_S	*N*	B_S	*N*	B_S	*N*	B_S	*N*	B_S	*N*
COI-1	1.00	49	*	-	1.00	16	*	-	1.00	11	*	-
COI-2	0.97	2113	*	-	0.96	538	1.00	76	0.97	311	0.98	235
COI-3	0.96	650	*	-	0.94	326	1.00	33	0.96	159	1.00	75
MCB	0.95	1426	*	-	0.88	203	*	-	0.95	841	0.97	364
cytM	0.90	935	*	-	0.80	177	*	-	0.99	272	0.94	476
16Sr	0.98	1730	*	-	0.97	624	1.00	118	0.99	243	0.99	560
16Sr2	0.98	1769	*	-	0.97	624	1.00	118	0.99	286	0.99	560
16Smam	0.83	3242	*	-	0.83	518	0.76	1297	0.90	351	0.90	1063

In the analysis of Resolutions, only primers amplifying more than 10 species per taxon are considered.

N: number of sequences amplified from Genbank.

* The resolution was not calculated as the primer pairs amplified 10 or less different species for this taxon.

of DNA barcoding, and can help the broad scale analysis of life on earth. However, some authors argued that it is impossible that a single short sequence will be enough to distinguish all members of all species [12]. In this context, explicit *in silico* approaches like the one presented in this study allow analysing the properties of different sets of primers, and identifying the most appropriate ones *a priori*.

In silico vs. real PCR

The real *in vitro* amplification pattern depends on PCR conditions. Controlling the PCR conditions can alter amplification results, and thus the taxonomic coverage of primers. For example, low annealing temperature and high concentration of $MgCl_2$ reduce the specificity of primers in real-world PCR, and can thus allow amplification of target sequences with a larger number of mismatches in the primer regions [17]. Our *in silico* analyses have been performed allowing two mismatches. These parameters correspond well to actual amplification at rather high annealing temperatures (Figure 2), in accordance with previously published environmental genetics studies [22]. Nevertheless, these stringent conditions probably lead us to predict more false negative results (non electronic amplification of amplifiable sequences) than false positive ones (electronic amplification of non amplifiable sequences). Increasing the authorized mismatches can simulate more relaxed conditions, but the strict relationship between electronic and experimental conditions cannot be formally described. On the other hand, stringent PCR conditions reduce the risk of amplifying unwanted regions of the genome (see below), particularly when using degenerate primers. Furthermore, our study focused on *sensu lato* barcode primer pairs. These studies often amplify DNA extracted from environmental samples, which may represent a mix of the DNA of several taxa [6]. Considering this, primers and PCR conditions must be as specific as possible, because the rare species with a low number of mismatches in the primer region (Figure 4) are expected to be overamplified and overrepresented in the PCR products, while species that are present, but with a higher number of mismatches, may not be amplified enough to yield sequences. Therefore, "ideal" primers would have a constantly low number of mismatches, leading to a less biased estimate of species presence.

EcoPCR can also be used to simulate less stringent PCR conditions, allowing more mismatches. With this approach, primers can amplify a much larger number of species (Figure 3). For example, in our stringent *in silico* analysis the primers Uni-Minibar showed limited taxonomic coverage, and amplified very few vertebrates (table 2). Conversely, the PCRs performed by Meusnier *et al.* [20] showed that these primers can amplify nearly 100% of fish and Amphibians, at an annealing temperature of 46°C. Results coherent with Meusnier *et al.* [20] can be obtained using ecoPCR by allowing a large number of mismatches (up to eight) (Figure 4). Taking into account all these considerations, we have to assume that the taxonomic coverage B_c estimated from ecoPCR is not an exact value, but it reflects the relative capacity of primer pairs to amplify a broad variety of taxa. For example, the fact that 16Sr amplifies a much larger number of species of amphibians than COI-2H [[11,23], see also [24] for a different approach] was correctly predicted by *in silico* analyses (see Figure 2, table 2).

Pseudogenes are a further potential issue in barcoding analysis; our approach may be affected by this trap. For instance, in our analyses none of the primers amplified nuclear sequences. However, nuclear sequences are underrepresented in Genbank; furthermore, the *in silico* amplification of pseudogenes would require the presence of a target nuclear sequence and both the corresponding primer regions, i.e., a good coverage of nuclear genome. Therefore it is difficult that ecoPCR hits nuclear pseudogenes, which can nevertheless be amplified by *in vitro* PCR, particularly under relaxed (e.g., low annealing temperature) conditions. Another potential issue of our approach is that the adjoining primer regions of sequences submitted to the databases are not a queryable portion of the database, therefore limiting the number of sequences obtained when ecoPCR is run over the

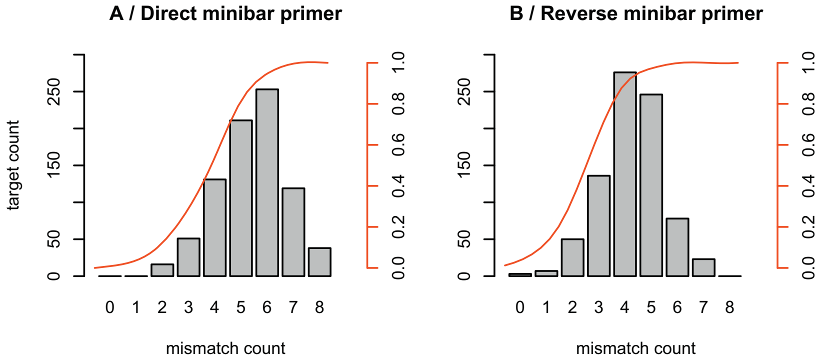


Figure 4 Mismatches between Uni-Minibar primers and vertebrate sequences. The histograms show the distribution of mismatch counts between (A) direct Uni-Minibar primer or (B) reverse Uni-Minibar primer and their target loci on mitochondrial DNA, as revealed by a ecoPCR run using Uni-Minibar primers to amplify our mitochondrial reference database. For this run, 8 mismatches were tolerated for each primer. The red curve and the associated right axis represent the cumulative fraction of amplified sequences with less that *m* mismatch. We can observe than a few species present a small count of mismatches; these sequences with a few mismatches are expected to be advantageously amplified in a DNA mix containing multiple species.

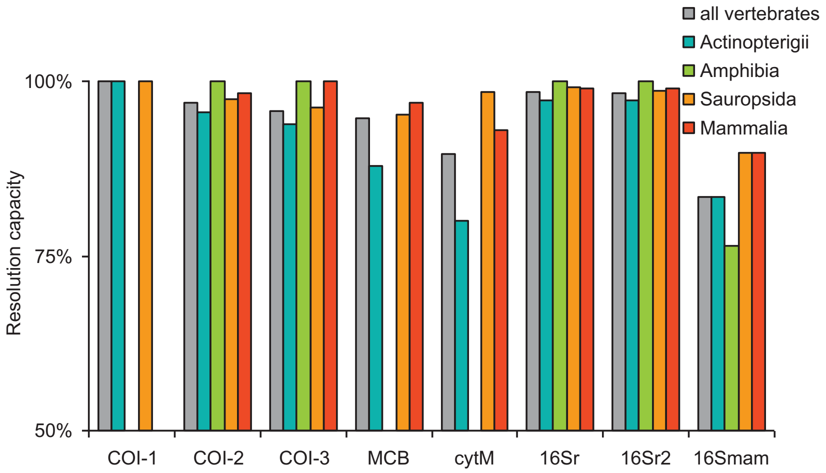


Figure 5 Resolution capacity of barcodes tested over the entire GenBank. Resolution is reported only for primer × taxon combinations that amplified more than 10 species. In all cases, resolution was > 50%.

entire GenBank. To partially address this issue, the assessment of taxonomic coverage was performed on species for which the whole mitochondrial genome was available, and therefore both target sequences and flanking regions are present. The increasing availability of whole mitochondrial genomes due the improvement sequencing technologies, and the rising of phylogenomics may reduce this limitation in the next future.

The correspondence between *in silico* and real PCR is certainly more accurate for the resolution capacity, still potential sources of bias remain. Our approach is based on the analysis of all the sequences deposited in GenBank, i.e., including thousands of vertebrate species in the example developed here. Assuming that all GenBank sequences are assigned to the correct species in the database, such approach uses the same kind of information than large scale barcoding studies. Clearly, the availability of sequences in different clades depends on the previous use of markers. For example, GenBank includes a very large number of COI sequences for Actinopterigii, while most of the mitochondrial sequences of mammals and amphibians are 16S. Furthermore, annotation errors are present in Genbank [25], and the error rate might be clade dependent. The B_S index is sensible to these errors, leading to an underestimation of B_S; therefore, as for B_C previously, B_S should be considered as a relative measure of primer performance.

Comparison of vertebrate barcodes

Universality is a key feature of barcodes, and several strategies exist that can increase the taxonomic coverage of primer pairs. One strategy consists in making cocktails of degenerate primers. For example, the COI-2 primer pair [18] had one of the highest taxonomic coverages (figure 2). A predictable drawback of degenerate primers is a limited specificity with regards to the target DNA sequence amplified. However, our *in silico* PCRs performed on the whole GenBank did not amplify incorrect regions. All sequences amplified by the COI-2 primer pair were labelled in GenBank as mitochondrial COI, suggesting that these primers maintained enough specificity.

An alternative strategy consists in designing universal primers on highly conserved regions. This strategy has been used for example on the 16S, that exhibits some highly conserved regions in vertebrates [26]. The primers amplifying the 16S [[27,28]; this study] were very powerful, and had the highest taxonomic coverage and resolution capacity in vertebrates (Figure 2, Figure 3, table 2). The 16S region has been investigated as an alternate barcode locus for amphibians [11] but COI has not been rejected [24]. Some studies advocated that 16S has a too low rate of molecular evolution, and thus does not hold enough interspecific variation for a correct species identification [1]. Our analysis suggests that, at least in vertebrates, 16S has the same resolution capacity as COI, when using sequences with comparable length (500-600 bp), and therefore can be a good candidate site for barcoding. Nevertheless, the good performance of 16S observed in vertebrates may not be valid in other taxa; our *in silico* approach can be a key tool to analyse this possibility.

Long barcodes (500-600 bp) like the standard COI and 16S barcodes have a high resolution capacity, and are ideal candidates, for example, to unambiguously identify taxa in the context of the original DNA barcoding usage. However, studies analysing environmental samples or degraded DNA require the use of shorter DNA fragments [6,7,13,20,22,29] even though those smaller regions include less information. We have included in our analysis two primer pairs amplifying short sequences that can be used for such analyses: Uni-Minibar [20] and 16Smam [21], which amplify sequences of 130-140 bp. Our analysis did not amplify enough sequences to evaluate the overall performance of Uni-Minibar, but allowed estimating the taxonomic coverage of 16Smam, which was very high for mammals (i.e., the taxon for which the primers have been designed), and lower for the other clades (Figure 2). This short barcode had the lowest resolution capacity for identification at the species level (Figure 3). However, in many cases species identification is not needed in ecological barcoding, as information on the genus or family can be already valuable [6,7,13,29]. Indeed, the resolution of 16Smam was much higher if the aim was the identification at the genus or family level (resolution capacity of 96% and 100%, respectively; results not shown).

Our analysis focused on vertebrates, because several primers have been proposed for their *sensu lato* barcoding. Furthermore, the *in silico* assessment of primers strongly depends on the sequences in online databases; vertebrates are the phylum best covered by available sequences, therefore they are the ideal focus of a methodological analysis. Nevertheless, biodiversity on Earth is dominated by other phyla, such as arthropods and molluscs: The evaluation method describe here can be applied to these taxa and to any other ones, considering that the precision of the estimated B_S and B_C indices is directly linked to the amount and the quality of available sequences in public database corresponding to the studied clade.

Conclusion

Based on our *in silico* analyses, the different barcodes tested showed dissimilar adequacy to be used according to the five clades of vertebrates studied. If we consider all possible applications of *sensu lato* barcoding, no single barcode could be identified as the best for all vertebrates. The primers amplifying COI-2 showed the highest taxonomic coverage in Actinopterigii and Sauropsida, while those amplifying 16Sr/16Sr2 showed the highest coverage of Amphibians and Mammals (Figure 3, table 2). Furthermore, the barcodes with the highest taxonomic coverage and resolution capacity (i.e., COI-2, 16Sr, 16Sr2) amplified long fragments, which can make their application problematic for describing biodiversity within environmental samples. In such a context, it is useful to select *a priori* the barcode that best suited the research topic. Our *in silico* method can help identifying the most appropriate barcode according to different aims. Such formal approach, which is possible thanks to the availability of bioinformatics tools and large public databases, can focus on target taxa or DNA regions and would make easier the validation of new barcodes by reducing the number of candidate primer pairs to be tested *in vitro*.

Acknowledgements

We thank three anonymous reviewers for constructive comments on an earlier version of the manuscript. GFF was funded by CNRS and LECA; this work has been partly supported by the Alocad project (ANR-06-PNRA-004-02) and by the European Project EcoChange (FP6-036866).

Author details

[1]Laboratoire d'Ecologie Alpine, CNRS UMR 5553, Université Joseph Fourier, BP 53, F-38041 Grenoble Cedex 9, France. [2]Dipartimento di Biologia, Università degli Studi di Milano. Via Celoria 26, 20133 Milano Italy. [3]Dipartimento di Scienze dell'Ambiente e del Territorio, Università degli Studi di Milano Bicocca. Piazza della Scienza 1, 20126 Milano Italy.

Authors' contributions

GFF, EC, PT and FP participated to the design of the study; JB and EC developed ecoPcr; TR and EC developed Bs and Bc indices, WS performed *in vitro* experiments; GFF and SZ performed the analyses; GFF, EC, PT and FP wrote the paper. All authors read and approved the final manuscript.

Received: 6 November 2009 Accepted: 16 July 2010
Published: 16 July 2010

References

1. Hebert PDN, Cywinska A, Ball SL, DeWaard JR: **Biological identifications through DNA barcodes.** *Proc R Soc B* 2003, **270**:313-321.
2. Barber P, Boyce SL: **Estimating diversity of Indo-Pacific coral reef stomatopods through DNA barcoding of stomatopod larvae.** *Proc R Soc B* 2006, **273**:2053-2061.
3. Hebert PDN, Penton EH, Burns JM, Janzen DH, Hallwachs W: **Ten species in one: DNA barcoding reveals cryptic species in the neotropical skipper butterfly Astraptes fulgerator.** *Proc Natl Acad Sci USA* 2004, **101**:14812-14817.
4. Janzen DH, Hallwachs W, Blandin P, Burns JM, Cadiou JM, Chacon I, Dapkey T, Deans AR, Epstein ME, Espinoza B, *et al*: **Integration of DNA barcoding into an ongoing inventory of complex tropical biodiversity.** *Mol Ecol Resour* 2009, **9**:1-26.
5. Willerslev E, Hansen AJ, Binladen J, Brand TB, Gilbert MTP, Shapiro B, Bunce M, Wiuf C, Gilichinsky DA, Cooper A: **Diverse plant and animal genetic records from Holocene and Pleistocene sediments.** *Science* 2003, **300**:791-795.
6. Valentini A, Pompanon F, Taberlet P: **DNA barcoding for ecologists.** *Trends Ecol Evol* 2009, **24**:110-117.
7. Valentini A, Miquel C, Nawaz N, Bellemain E, Coissac E, Pompanon F, Gielly L, Cruaud C, Nascetti G, Wincker P, *et al*: **New perspectives in diet analysis based on DNA barcoding and parallel pyrosequencing: the *trn*L approach.** *Mol Ecol Resour* 2009, **9**:51-60.
8. Hebert PDN, Ratnasingham S, deWaard JR: **Barcoding animal life: cytochrome c oxidase subunit 1 divergences among closely related species.** *Proc R Soc B* 2003, **270**:S96-S99.
9. Kress WJ, Wurdack KJ, Zimmer EA, Weigt LA, Janzen DH: **Use of DNA barcodes to identify flowering plants.** *Proc Natl Acad Sci USA* 2005, **102**:8369-8374.
10. Hollingsworth PM, Forrest LL, Spouge JL, Hajibabaei M, Ratnasingham S, Bank van der M, Chase MW, Cowan RS, Erickson DL, Fazekas AJ, *et al*: **A DNA barcode for land plants.** *Proc Natl Acad Sci USA* 2009, **106**:12794-12797.
11. Vences M, Thomas M, Bonett RM, Vieites DR: **Deciphering amphibian diversity through DNA barcoding: chances and challenges.** *Phil Trans R Soc B* 2005, **360**:1859-1868.
12. Rubinoff D, Cameron S, Will K: **Are plant DNA barcodes a search for the Holy Grail?** *Trends Ecol Evol* 2006, **21**:1-2.

13. Deagle BE, Kirkwood R, Jarman SN: **Analysis of Australian fur seal diet by pyrosequencing prey DNA in faeces.** *Mol Ecol* 2009, **18**:2022-2038.
14. Moritz C, Cicero C: **DNA Barcoding: Promise and pitfalls.** *PLoS Biol* 2004, **2**:1529-1531.
15. Altschul SF, Madden TL, Schaffer AA, Zhang JH, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res* 1997, **25**:3389-3402.
16. Wu S, Manber U: **Agrep - a fast approximate pattern-matching tool.** *Proceedings of the Winter 1992 USENIX Conference San Francisco USA 20-24 Jan 1992* Berkeley, USA 1992, 153-162.
17. Elrich HA, Gelfand D, Sninsky JJ: **Recent advances in the Polymerase Chain Reaction.** *Science* 1991, **252**:1643-1651.
18. Ivanova NV, Zemlak TS, Hanner RH, Hebert PDN: **Universal primer cocktails for fish DNA barcoding.** *Mol Ecol Notes* 2007, **7**:544-548.
19. Hajibabaei M, Smith MA, Janzen DH, Rodriguez JJ, Whitfield JB, Hebert PDN: **A minimalist barcode can identify a specimen whose DNA is degraded.** *Mol Ecol Notes* 2006, **6**:959-964.
20. Meusnier I, Singer GAC, Landry J-F, Hickey DA, Hebert PDN, Hajibabaei M: **A universal DNA mini-barcode for biodiversity analysis.** *BMC Genomics* 2008, **9**:214.
21. Taylor PG: **Reproducibility of ancient DNA sequences from extinct pleistocene fauna.** *Mol Biol Evol* 1996, **13**:283-285.
22. Ficetola GF, Miaud C, Pompanon F, Taberlet P: **Species detection using environmental DNA from water samples.** *Biol Lett* 2008, **4**:423-425.
23. Vences M, Thomas M, Meijden van der A, Chiari Y, Vieites DR: **Comparative performance of the 16S rRNA gene in DNA barcoding of amphibians.** *Front Zool* 2005, **2**:5.
24. Smith MA, Poyarkov NA, Hebert PDN: **CO1 DNA barcoding amphibians: take the chance, meet the challenge.** *Mol Ecol Resour* 2008, **8**:235-246.
25. Schnoes AM, Brown SD, Dodevski I, Babbitt PC: **Annotation error in public databases: Misannotation of molecular function in enzyme superfamilies.** *PLoS Comput Biol* 2009, **5**.
26. Kocher TD, Thomas WK, Meyer A, Edwards SV, Paabo S, Villablanca FX, Wilson AC: **Dynamics of mitochondrial-DNA evolution in animals - Amplification and sequencing with conserved primers.** *Proc Natl Acad Sci USA* 1989, **86**:6196-6200.
27. Palumbi SR: **Nucleic acids II: the polymerase chain reaction.** *Molecular Systematics* Sunderland, Massachusetts: Sinauer & Associates;Hills MD, Moritz C, Mable BK 1996, 205-247.
28. Palumbi SR, Martin A, Romano S, McMillan WO, Stice L, Grabowski G: *The simple fool's guide to PCR, ver 2* Honolulu: University of Hawaii 1991.
29. Pegard A, Miquel C, Valentini A, Coissac E, Bouvier F, Francois D, Taberlet P, Engel E, Pompanon F: **Universal DNA-Based Methods for Assessing the Diet of Grazing Livestock and Wildlife from Feces.** *J Agric Food Chem* 2009, **57**:5700-5706.
30. Verma SK, Singh L: **Novel universal primers establish identity of an enormous number of animal species for forensic application.** *Mol Ecol Notes* 2003, **3**:28-31.
31. Meyer R, Hoffelein C, Candrian U: **Polymerase chain reaction restriction fragment length polymorphism analysis: A simple method for species identification in food.** *Journal of AOAC International* 1995, **78**:1542-1551.

doi:10.1186/1471-2164-11-434
Cite this article as: Ficetola *et al.*: An *In silico* approach for the evaluation of DNA barcodes. *BMC Genomics* 2010 **11**:434.

2.3 Complete Formalization Of B_c And B_s

In the article above the complete mathematical formalization of B_c and B_s indices is missing. It is detailed in this section. For this purpose we need to define some sets and relations. As shown in the Figure 1 of publication above, we define following sets:

$\mathcal{T} = \{t_i\}$	The set of all taxa.
$\mathcal{I} = \{id_i\}$	The set of all individuals.
$\mathcal{B} = \{b_i\}$	The set of all barcode sequences.
$\mathcal{R} = \{r_i\}$	The set of all barcode regions.
$\mathcal{L} = \{l_i\}$	The set of all taxonomic levels (ranks).

And we define following relations on these sets:

$E : \mathcal{T} \mapsto \mathcal{I}$	Membership relation of an individual to a taxon.
$E_L : \mathcal{L} \mapsto \mathcal{T}$	Membership relation of a taxon to a taxonomic level.
$E' : \mathcal{R} \mapsto \mathcal{B}$	Membership relation of a barcode to a region.
$Img : \mathcal{I} \mapsto \mathcal{B}$	Gives barcodes identifying an individual.

The set of all taxa amplified by the region r detectable by the primer pair defining this region are given by:

$$\beta(r) \equiv E^{-1}(Img^{-1}(E'(r))). \tag{2.3.1}$$

Since $E_L(l)$ gives the set of taxa belonging to a taxonomic level l, so finally we denote taxa of this taxonomic level amplified by the region r as:

$$\alpha(r, l) \equiv \beta(r) \cap E_L(l). \tag{2.3.2}$$

2.3.1 Complete Formalization Of B_c

The coverage index as defined in above article is the ratio of total number of amplified taxa to the total number of taxa of the same taxonomic level in the input data set. The computation of this index is only possible if the taxonomic content of the data set is fully defined.

From the above relations we define $B_c : \mathcal{R} \times \mathcal{L} \mapsto \mathbb{R}$ the fraction of taxa of a taxonomic level l detectable by the primer pair defining the region r.

$$B_c(r,l) = \frac{|\alpha(r,l)|}{|E_L(l)|} \,. \tag{2.3.3}$$

Following definition 2.3.3, identifying the best region in term of coverage corresponds to problem 1

Problem 1.

$$\textit{find } r \textit{ as } \quad B_c(r,l) \textit{ is } \max .$$

2.3.2 Complete Formalization Of B_s

In the above publication, barcode *specificity* (B_s) is defined as the ability of a region to discriminate between two taxa, or the ability of a region to unambiguously identify a taxon. We further said that a taxon is unambiguously identified if it owns a barcode region that is not shared by any other taxa of the same taxonomic rank. In order to compute the number of unambiguously identified taxa, we need to define some more relations.

Using above sets and relations, we define:

$$\Omega(t,r) = Img(E(t)) \cap E'(r), \tag{2.3.4}$$

where Ω gives us the set of all barcodes of a region r identifying individuals of a taxon t. And inversely the set of all individuals (may belong to multiple taxa) identified by a barcode of region r is given as:

$$Img^{-1}(\Omega) = \bigcup_i Img^{-1}(b_i \mid b_i \in \Omega). \tag{2.3.5}$$

We said that a taxon t is unambiguously identified (or well identified) by a barcode region r if and only if

$$Img^{-1}(\Omega(t,r)) = E(t). \tag{2.3.6}$$

If we denote the above set of well identified taxa by ϵ as:

$$\epsilon \equiv \{t \mid \textit{equation } 2.3.6 \textit{ holds}\}, \tag{2.3.7}$$

then the specificity B_s of a region r for a taxonomic level l is given as:

$$B_s \equiv \frac{|\{t \mid t \in \epsilon\}|}{|\alpha(r,l)|}. \tag{2.3.8}$$

Following this definition, identifying the best region in term of specificity corresponds to problem 2

Problem 2.

$$\textit{find } r \textit{ as } \quad B_s(r,l) \textit{ is } \max .$$

2.3.3 Extending The Definition Of B_s

The strict equality between left and right sides of equation 2.3.6 gives rise to a potential problem of falsely decreasing value of B_s. Looking at Figure 1 of article, we can see that taxa T_2 and T_3 are ambiguous because barcode sequence b_4 is shared between individuals of these taxa. This reduces the specificity value to $1/3$ because only 1 taxon is well-identified out of 3. There may be two potential reasons for individual I_6 to own b_4: first, this individual shares its barcode sequences with other taxa, rendering both the taxa sharing the same barcode, as not well-identified. But a second hypothesis that has to be considered is, since public data bases like *Genbank* contain many errors in taxonomical annotation, it is quite possible that this individual I_6 actually belongs to the other taxa T_2. This misassigned taxon T_3 makes T_2 ambiguous. This second hypothesis results in a decreased value of barcode $specificity$. In order to tackle this problem and not to falsely decrease the specificity we can extend the definition of barcode specificity to allow some errors in annotation. We say that a taxon t is identified by a barcode region r allowing a Q false positive errors rate if and only if

$$\left.\begin{array}{r} E(t) \subseteq Img^{-1}(\Omega(t,r)) \\ \text{and} \quad |Img^{-1}(\Omega(t,r)) \cap \bar{E}(t)| \leqslant Q\ |Img^{-1}(\Omega(t,r))| \end{array}\right\} \mathcal{E}_Q(t,r). \tag{2.3.9}$$

This defines a mapping $\mathcal{E}_Q$ from $\mathcal{T}$ to $\mathcal{R}$. This mapping has two conditions: i) $E(t) \subseteq Img^{-1}(\Omega(t,r))$, which means that the barcodes of region r identifying the individuals of taxon t may also identify individuals of some other taxa. ii) $|Img^{-1}(\Omega(t,r)) \cap \bar{E}(t)| \leqslant Q\ |Img^{-1}(\Omega(t,r))|$, this condition means that the number of individuals identified by barcodes of region r not belonging to taxon t are not more than Q percent of the total individuals identified by r. If these two conditions hold then extended definition of B_s is given as:

$$B_s(r,l,Q) \equiv \frac{|\{t \mid t\ \in \mathcal{E}_Q(t,r)\}|}{|\alpha(r,l)|}. \tag{2.3.10}$$

We can observe that the equation 2.3.10 is equivalent to the equation 2.3.8 if $Q = 0$. The result of this relaxed definition is an increase in B_s value.

The main problem of using this new version of B_s is that for precise taxonomic range like species, the number of sequences belonging to each taxa is low on average. For example, if we consider two species, sp_1 and sp_2, each of them represented by 2 sequences s_a, s_b and s_c, s_d respectively; and s_d is erroneously annotated as belonging sp_1. In this case we need to set $Q > 1/3$ to tackle this error. But this high value of Q is unrealistic, and could lead to artificially increased value of B_c. A solution to this problem could be, not to consider each decision (this taxa is unambiguously identified) individually but as a set of decisions noised by a binomial process of wrong annotation of parameter p and n, where $p =$ error rate in *Genbank* $\sim$ 10%. Under this hypothesis we would have to select the set of decisions maximizing the likelihood and then compute B_s according to it.

2.3.4 Falsely Increased Value of B_s

A taxon owns a set of barcode sequences that belong to a barcode region. According to our definition, an unambiguously identified taxon shares none of its barcode sequences with another taxa. Two taxa are considered to be sharing a barcode sequence if at least one barcode sequence of the first taxon is strictly identical to a sequence included in the set of barcode sequences of the second taxon. If we consider the possibility of errors during sequencing or PCR amplification, then it is possible to have certain taxa sharing some barcode sequences. Given two taxa t_1 and t_2 with one barcode sequences each *i.e.* s_1 and s_2 respectively, if s_1 and s_2 differ by only one base pair we will not be able to distinguish them during the analysis of the results. If s_1 and s_2 are present in the results, we can propose three possibilities : i) Both t_1 and t_2 are actually present in the sample, ii) only t_1 is present and s_2 is a reading error of s_1, iii) the opposite situation. We can deal with this problem by changing the initial definition as following : Two taxa t_1 and t_2 and their associated sets of barcode sequences s_1 and s_2 respectively are considered as unambiguously identified if and only if

$$\forall\ s_i \in s_1 \text{ and } s_j \in s_2 : \min(d_H(s_i, s_j)) > d_{min} \tag{2.3.11}$$

If $d_{min} = 0$ this new definition is identical to the original one. By increasing d_{min} we will have a measure of B_s more robust but with a smaller value.

For computing B_s following this new definition, we build a graph $G(S, D)$ where S, the set of vertices, is composed of all possible barcode sequences s for the considered marker and D is a relation defined as $d_H(s_i, s_j) \leq d_{min}$. Each $c \in \mathcal{C}$ the set of all connected component composing G can be considered as an equivalence class of barcode sequences. Thus B_s can be computed by substituting the set $\mathcal{B}$ with $\mathcal{C}$ in the original definition.

2.4 Conclusion

In this chapter we have given detailed formalization of two measures i.e B_c and B_s. These two indices are extremely helpful for evaluating the quality of barcode regions for a given taxonomic rank. Ecologists can take advantage of huge amount of data available due to next generation sequencing techniques, and design barcode markers suitable for a particular study. In this context these two indices can be helpful for ranking the inferred barcode markers and for selecting the best markers, thus limiting the number of markers actually to be used in an experiment. Finally we have proposed two extensions to the definition of B_s due to the presence of errors in sequences. These extensions are not present in the publication as the article was published before we started working on the problem of errors (see chap 4, page 107).

2.5 Résumé

Ce chapitre traite de la comparaison objective des marqueurs utilisés pour le DNA barcoding. Ce chapitre articulé autours d'une de mes publications présente deux indices quantitatifs et formels développés durant ma thèse pour mesurer la qualité d'un marqueur. L'application de PCR "in silico" nommée *ecoPCR* et utilisé pour le calcule de ces indices y est également présenter. Ce chapitre s'achève sur la présentation de quelques extensions théoriques pour ces deux indices permettent de considérer les erreurs de séquences et d'annotation taxonmiques.

CHAPTER 3

Optimal Primer Design

In the recent decade, DNA barcoding has become a method of choice for characterizing species diversity. The method has become equally popular among taxonomists and ecologist given that morphological identification is not always possible and moreover access to the species of interest may not always be feasible. This is the case for measuring the microbial diversity or determining the diet of carnivores, where getting the stomach content is difficult. However in this case, feces samples can be easily collected, DNA contained in them can amplified by PCR and the resulting amplicons can be sequenced to determine the diet. For such applications and for many others, DNA barcoding has been proven successful.

However one major challenge in DNA barcoding is the design of optimal barcode markers suitable for metabarcoding applications particularly. It has already been discussed in the introduction of this thesis that standard markers are available for classical DNA barcoding applications but no standard markers exist for metabarcoding applications. Almost all of the available primer design programs work for small number of short sequences. Our objective is to be able to infer barcode markers by scanning full genomes and looking for markers from huge databases of sequences in order to design highly conserved primer pairs and universal barcode markers. None of the available programs are efficient enough to be able to run on sequence of more than several Megabytes. In this chapter I detail my work on the design of optimal barcode design process.

I have designed the program *ecoPrimers* which is highly efficient being able to successfully scan the fully sequenced bacterial genomes and design optimal barcode markers. The barcode markers designed are optimized using the B_c and B_s quality indices described in chapter 2. The article on program *ecoPrimers* is accepted in the journal of *Nucleic Acid Research*. The publication follows on the next page.

Published online 19 September 2011 *Nucleic Acids Research, 2011, Vol. 39, No. 21* ***e145***
doi:10.1093/nar/gkr732

ecoPrimers: inference of new DNA barcode markers from whole genome sequence analysis

Tiayyba Riaz[1], Wasim Shehzad[1], Alain Viari[2], François Pompanon[1], Pierre Taberlet[1] and Eric Coissac[1,]*

[1]Laboratoire d'Ecologie Alpine (LECA) CNRS UMR 5553 2233, Université Joseph Fourrier, BP 53, 38041 Grenoble Cedex-9 and [2]INRIA Rhône-Alpes – Projet Bamboo, ZIRST-655 Avenue de l'Europe, 38334 Montbonnot Cedex, France

Received June 9, 2011; Revised and Accepted August 23, 2011

ABSTRACT

Using non-conventional markers, DNA metabarcoding allows biodiversity assessment from complex substrates. In this article, we present *ecoPrimers*, a software for identifying new barcode markers and their associated PCR primers. *ecoPrimers* scans whole genomes to find such markers without *a priori* knowledge. *ecoPrimers* optimizes two quality indices measuring taxonomical range and discrimination to select the most efficient markers from a set of reference sequences, according to specific experimental constraints such as marker length or specifically targeted taxa. The key step of the algorithm is the identification of conserved regions among reference sequences for anchoring primers. We propose an efficient algorithm based on data mining, that allows the analysis of huge sets of sequences. We evaluate the efficiency of *ecoPrimers* by running it on three different sequence sets: mitochondrial, chloroplast and bacterial genomes. Identified barcode markers correspond either to barcode regions already in use for plants or animals, or to new potential barcodes. Results from empirical experiments carried out on a promising new barcode for analyzing vertebrate diversity fully agree with expectations based on bioinformatics analysis. These tests demonstrate the efficiency of *ecoPrimers* for inferring new barcodes fitting with diverse experimental contexts. *ecoPrimers* is available as an open source project at: http://www.grenoble.prabi.fr/trac/ecoPrimers.

INTRODUCTION

DNA barcoding opens new opportunities for biodiversity research. This technique is now considered to be a powerful tool, both for taxonomical (1) and ecological (2) studies. Taxonomies based solely on morphological analyses are sometimes problematic due to either convergence in phenotypes among distantly related species, or the failure to identify cryptic species where morphologic divergence has not kept pace with genetic divergence (3). Though the original aim of DNA barcoding was to assign an unambiguous molecular identifier to each taxon (1), today new DNA barcoding applications are emerging. These applications apply DNA barcodes not as a means to unambiguously identify a single specimen from a taxonomical point of view, but as a tool for better characterizing a set of taxa from a complex biological sample. This metabarcoding approach (i.e. the simultaneous identification of many taxa from the same sample) has a wide range of applications in forensics, ecology and palaeoecology.

Following the original (*sensu stricto*) barcode definition, a barcode marker must be as universal as possible and must contain enough information to discriminate between closely related species and to discover new ones. The Consortium for the Barcode of Life (CBoL: http://www.barcodeoflife.org) leads the standardization of such markers. For example, the *COI* gene is recommended for animal barcoding (1). However, in ecological research, other constraints must sometimes be considered when selecting a barcode marker and its associated primers. As a consequence, the standardized *COI* animal barcode that clearly fulfills all the requirements for specimen identification (1) is not always the most efficient one for a metabarcoding approach.

Metabarcoding constraints on the locus choice

Sensu stricto barcode applications prefer long barcode markers with high discrimination capacity and, if possible, high phylogenetic information content. For these reasons the *COI* gene for animals (1) and *rbc*L and *mat*K genes for plants (4) are recommended by CBoL. Metabarcoding has a different aim and requires different

*To whom correspondence should be addressed. Tel: (+33) 4 76 63 54 50; Fax: (+33) 4 76 51 42 79; Email: eric.coissac@inrialpes.fr optimality criteria for the markers employed: (i) as the DNA will often be degraded (and to minimize the risk of chimeric sequences) shorter amplicons are needed, and (ii) to minimize amplification biases in mixed-template reactions, the primers need to be highly conserved. Furthermore, taxonomic resolution at the species level is not always required. Identification at a higher taxonomic level (e.g. family, order, etc.) is sometimes sufficient. Thus in some conditions, it might be necessary to select a short marker even if its resolution is low.

Metabarcoding constraints on the primer choice

Sensu stricto barcode applications usually rely on PCR amplifications from good quality DNA extracted from a single specimen. This allows the use of degenerate primers and relaxed PCR conditions, with the key constraint of amplifying the same highly informative standard locus from the broadest range of organisms. *A contrario*, metabarcode applications require robust PCR conditions allowing unbiased amplifications from a mix of several DNA templates which are often degraded [DNA extracted from modern and ancient soils (5,6), water (7) or animal feces (8,9)]. This imposes the use of highly conserved primers for simplifying PCR amplification conditions and reducing disequilibrium in amplification among the different DNA templates. Moreover, it can be advantageous to select primers amplifying only a subset of taxa for solving a given biological question (i.e. excluding the amplification of other taxonomic groups).

Tracking the ideal barcode markers

Ideal metabarcode markers should be short, highly discriminant, restricted to the studied clades and have highly conserved primer sites. Such ideal markers might not be the same among studies. In many cases this requires a specific pair of primers be designed to exactly fit the biological question.

The traditional method for identifying barcode regions is human observation of sequence alignments to locate two conserved regions flanking a variable one. This manual approach obliges barcode designers to work on well-known sets of genes. Based on this approach, several manually discovered barcode loci are in routine use today, including regions of protein encoding genes such as *COI* (1,12), *rbc*L or *mat*K (4), RNA genes like mitochondrial *12S* (13) or *16S* (14) rDNA and non-coding chloroplast regions such as the *trn*L intron (15) or the intergenic *trn*H-*psb*A region (16). Several tools exist to help biologists during the primer design step, but they were not often developed for the context of DNA barcoding. Among them, Primer3 (17) and QPrimer (18) use a single training sequence and were clearly not developed for designing versatile primers. TmPrime (19) and UniPrimer (20) can work on a training set of short sequences (i.e. gene sequences), allowing the design of primers that amplify several homologous sequences. But these tools are not adapted for long sequences (i.e. whole genomes) and do not take into account the taxonomic discrimination capacity of the amplified sequence during the primer selection process. More interestingly, PrimerHunter (21) was developed to select highly specific primers for distinguishing virus subtypes, a typical *sensu lato* barcoding application. Unfortunately, its efficiency on large data sets of long sequences is problematic. We were unable to run it on a 13.7 MB (Megabyte) database corresponding to the full set of whole mitochondrial genomes extracted from GenBank. Finally, Amplicon (22) allows for selecting specific primers to a group of aligned sequences and excluding a counterexample data set. But, as Amplicon requires aligned sequences, it can only design primers from a set of short regions compatible with multi-alignment software capacity and so cannot be run with a whole-genome data set.

To efficiently infer new metabarcode markers, we developed a software, *ecoPrimers*, fulfilling the following prerequisites: (i) the ability to scan a large database of whole genomes allowing the selection of markers without *a priori* identification, (ii) the ability to select highly conserved primers among a training set of sequences (example sequences) and possibly not amplifying a counterexample set of sequences (iii) the ability to test an amplified region for its capacity to discriminate among taxa. For achieving these goals, we took advantage of two indices previously proposed to evaluate *in silico* the relative quality of barcode primers in the context of metabarcoding (10). The first index, B_c, estimates the coverage or taxonomical amplification range of a primer pair. The second, B_s, evaluates the taxonomical discrimination capacity of the amplified marker among the amplified taxa. These indices have been successfully used by Bellemain *et al.* (11) to demonstrate the importance of primer selection for metabarcoding studies of fungal communities. *ecoPrimers* selects primer pairs by optimizing these two indices. A special effort was made to ensure computational efficiency of the program, and this was tested on the one thousand bacterial genomes currently available in public databases.

Here we used *ecoPrimers* to design specific primer pairs for bacterial, chloroplast and mitochondrial genomes. Validation by empirical experiments of the primer pairs selected to identify the vertebrates confirms that *ecoPrimers* proposed specific and robust primer pairs for amplifying target sequences. *ecoPrimers* is available as an open source software at: http://www.grenoble.prabi.fr/trac/ecoPrimers.

MATERIALS AND METHODS

Problem formulation

We assume that all sequences are texts over the DNA alphabet $\{A, C, G, T\}$, and that the orientation of sequences is unknown. Given a set of example sequences E_s and an optional second set of counterexample sequences C_s, we want to identify highly conserved primers which are present in the largest possible subset of E_s and in the smallest subset of C_s. Highly conserved primers are defined as words of length l_p, (i) strictly present in at least Q_s sequences of E_s, (ii) present in at least Q_e sequences of E_s with no more than e mismatches (optionally we can impose that these errors are not located in the n last 3′ bases of the primers to be more realistic in subsequent empirical DNA amplification), (iii) not present in more than Q_x sequences of C_s. The same approximative matching conditions used for Q_e are applied to this quorum. By default Q_s is set to 70% of $|E_s|$, Q_e is set to 90% of $|E_s|$ and Q_x is set to 10% of $|C_s|$. Identified potential primers are then paired with respect to their locations and orientation to allow amplification of those DNA fragments that are within the size range specified by the user.

Algorithm

In a nutshell, our method consists of five steps: (i) finding strict primers (i.e. without mismatch) from E_s respecting Q_s; (ii) using these strict primers as models to find their non-strict occurrences (i.e. with mismatches) in E_s to check Q_e and in C_s to check Q_x; (iii) building the primer pairs, (iv) evaluating B_c and B_s indices to select the best primers, and (v) estimating the melting temperature of each of the primers in selected pairs.

Finding strict repeats. Finding conserved regions among a set of sequences is an equivalent problem to finding repeats among those sequences. Identification of repeats in DNA sequences is a well-known problem in bioinformatics and many efficient data structures and associated algorithms exist for finding strict repeats, such as KMR (23), suffix tree (24) and suffix array (25). These algorithms work well on short sequences but are not efficient enough for us in terms of memory usage for finding repeats in a quorum of a large number of very long sequences (i.e. the set of all whole sequenced bacterial genomes available in public databases, approximatively 1000 genomes and 3 Gb (gigabases) of sequences). The best implementation of suffix tree was developed in Reputer (26). It uses about 12.5 bytes per nucleotide to build the data structure. This compact implementation is based on a 32 bit architecture; consequently it cannot manipulate sequence data larger than 340 Mb (megabases). Similarly, the most compact implementation of KMR is done in RepSeek, (27) which uses about 9 bytes per nucleotide on a 32 bit architecture, corresponding to a limit of 475 Mb. The last structure, suffix array, requires 4 bytes per nucleotide on a 32 bit, and 4 more bytes to be efficiently used to infer repeats. These two values have to be multiplied by 2 on a 64 bit architecture. Finally, as we do not assume that all the sequences are in the same orientation, we have to encode the direct and the reverse strand in the data, multiplying by two the memory requirement.

These three algorithms simultaneously identify conserved motifs and the positions of their occurrences. Following our brief description of the *ecoPrimers* algorithm, we just need the motif and the number of the sequences in which they occur. We do not need their exact positions, as they will be recomputed in step (ii) taking into account mismatches. We take advantage of this to gain memory compactness.

For *ecoPrimers* we have developed a simple algorithm for finding strict repeats which is notably compact in memory. This algorithm is based on a sort and a merge algorithm and some data mining steps. The algorithm presented in Figure 1 (named Strict Primer Algorithm, SPA) gives the outline of our strict repeats finding procedure without a data mining step.

In the first step, we load all sequences in memory. Then we construct an empty list L_P that will contain the strict repeats found at the end of the algorithm as a set of couple (W, n) where W is a word and n is the number of sequences where it occurs. In the third step, for each input sequence S_i of E_s, we build L_W, the list of all overlapping words of length l_p. For purpose of compactness, words are saved as a 64-bit binary hash code (named further D_{code} or R_{code}) following the encoding schema $\{A = 00, C = 01, G = 10, T = 11\}$. This allows us to manipulate words up to 32 nucleotides long.

To look for repeats in both strands of a DNA sequence, standard algorithms are required to store direct and reverse sequences in their data structures. In a double stranded DNA sequence, occurrence position is defined by a position and an orientation. As in our algorithm, occurrence positions are not important at this stage, orientations of enumerated words do not have to be stored. Thus, if a word W occurs n times in both strands of a sequence, $\overleftarrow{W}$ the reverse complement corresponding word of W also occurs n times. Therefore we just need to count one of the two (W or $\overleftarrow{W}$). The actual counted word for a given word pair $(W, \overleftarrow{W})$ is the one corresponding to the smaller hash code between D_{code} and R_{code}.

Sorting (Step 7) is achieved using the Smoothsort algorithm (25,28). This algorithm has a complexity of $O(n\log n)$ in the worst case, as do several other sorting algorithms, but has a complexity near to $O(n)$ when the input array is almost ordered.

The merge (Step 9) of the two lists L_P and L_W is achieved in place and in a linear time using just an extra buffer of $size = minimum(|L_P|, |L_W|)$. During this merging step words that will not be able to respect Q_s are

```
1 - Load sequences in memory
2 - Create an empty pattern list $L_P$ of couples $(W,n)$ where
$W$ is a word and $n$ is the number of sequences where it
occurs;
for all sequences $S_i \in E_s$ do
    3 - Build empty list of binary words $L_W$;
    for all Words $W \in S_i$ of length $l_p$, do
        4 - Build $D_{code}$ the hash code of $W$;
        5 - Build $R_{code}$ the hash code of $\overleftarrow{W}$ the reverse
        complement of $W$;
        6 - Append $Minimum(D_{code}, R_{code})$ to $L_W$;
    end for
    7 - Sort $L_W$;
    8 - Remove duplicates;
    9 - Merge $L_W$ with $L_P$ updating count n in $L_P$;
    10 - Remove couple from $L_P$ that cannot meet $Q_s$
    conditions;
end for
```

Figure 1. Strict primer algorithm (SPA) used for finding strict repeats.

eliminated of L_P. Despite this, the $|L_P|$ increases quickly until $|E_s| - Q_s$ sequences are analyzed (Figure 2a). This technique is sufficient for data sets of reasonable size, but for large data sets like fully sequenced bacterial genomes having total size of approximately 3 Gb, it consumes a significant amount of memory. To overcome this problem a pre-filtration/data-mining step was added.

Data mining. Data mining used for finding strict repeats is based on the fact that all words W of size l_p present in at least Q_s sequences of E_s are composed only of words W_m of size $l_m \leq l_p$ present in at least Q_s sequences of E_s. Using the binary encoding schema presented previously, we built a complete hash table H_m of all words W_m of size $l_m = 13$. Each cell of this table stores the count of sequences where the corresponding word occurs. As we have $4^{13} = 67\,108\,864$ different words of size l_m, and for each word the hash table used 4 bytes, 256 MB of memory is required to store it. This size is small if we compare it to the 3 GB used to store the bacterial genome sequences and more than 8 GB used by SPA to store the L_P list corresponding to these sequences. H_m is built in a linear time.

To include data mining in SPA, we just added a condition on H_m in the building hash code methods of Steps 3 and 4 (Figure 1), verifying the assertion that no word $W_m \in W$ is present in less than Q_s sequences. As computation of the next hash code at Steps 3 and 4 is achieved by bit shifting of the previous one, only one lookup into H_m is required per hash code generated. Each lookup is done in constant time so data mining does not change the global complexity of the initial algorithm.

Finding approximate primers. In the above step we have found a list of words L_P which are present in at least Q_s of the E_s. In this step, we find the approximate occurrences of these words in all the example sequences $S_e \in E_s$ and all the counterexample sequences $S_c \in C_s$. For this purpose, we use these strict words as patterns and find their approximate occurrences using the *agrep* algorithm (29). At the end, we conserve only words occurring in more than Q_e sequences of E_s with no more than e errors (i.e. mismatches). From these words, the words which are not present in more than Q_x sequences of C_s are tagged as good primers.

Pairing the primers. Words must finally be paired to delimit potential barcode regions. Pairing is done for all the sequences with an almost linear time algorithm checking the minimal ($l_{\min}$) and maximal length ($l_{\max}$)

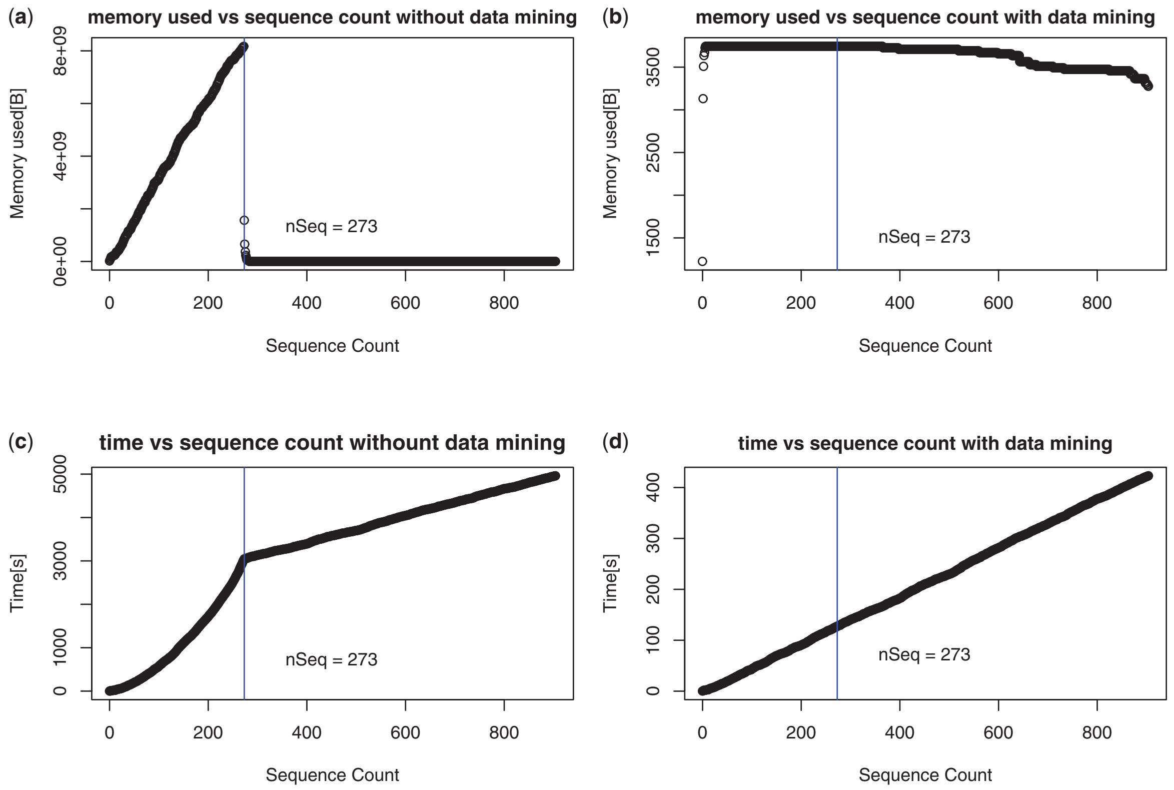


Figure 2. Comparison of time and memory usages of the both versions of the SPA. (**a**) Memory used with respect to the sequences processed without data mining step. Memory used increases rapidly until strict quorum (70%) starts taking effect after 271 (30% of 905) sequences have been processed (**b**) Same but with data mining step. Only a small number of prefix of 13 bases for primers of length18 bases pass the strict quorum, hence memory used is significantly small. (**c**) Time required to process the sequences without data mining increases exponentially until strict quorum starts making effect and after that time becomes linear. (**d**) With the data mining step added, time required becomes linear.

constraint imposed on the potentially amplified sequence. Each pair must contain at least one good primer (specificity of a single primer is enough to ensure specificity of the amplified region). A primer pairs is composed of two words and their relative orientation indicates which one of W and $\overleftarrow{W}$ must be used as primer. Once orientation is defined only pairs satisfying the constraint of no mismatches on the n last $3'$ bases of the primer are conserved.

Applying the quality indices. Once constructed, the primer pairs can be evaluated using both the indices B_c and B_s defined in Ficetola *et al.* (10). B_c the barcode coverage index is the ratio between the number of amplified taxa and $|E_s|$. B_s the barcode selectivity index is the ratio between the number of identified taxa and $|E_s|$. These indices can be efficiently computed in *ecoPrimers* using data stored during the pairing process.

Melting temperature calculation. ecoPrimers uses the nearest neighbor thermodynamic model (30) for melting temperature (*Tm*) computation. Using this technique we estimate *Tm* of the perfect match of the primer and of the worst match of the primer on the example sequence. The temperatures are calculated using the following formula:

$$T_m = \frac{\Delta H}{\Delta S + 0.368 \times N/2 \times \ln(Na^+) + R \times \ln(C)} \qquad \textbf{(1)}$$

Here, ΔH and ΔS are enthalpy and entropy changes for annealing reaction respectively. This annealing reaction results in a duplex having Watson–Crick base pairs. N is the total number of phosphates in the duplex, R is the universal gas constant, C is the total DNA concentration from (30) and Na^+ is the concentration of salt cations. ΔH and ΔS are computed by summing experimentally estimated contributions of constituting dimer duplexes as in (21).

Empirical *ecoPrimers* evaluation

ecoPrimers must be evaluated for its computational efficiency and the quality of its results. Efficiency was tested using the large *eubact* data set (*vide infra*). The quality of the results proposed by *ecoPrimers* can be checked by comparing proposed barcodes with ones currently used. If we assume that previously used barcodes were designed empirically but correctly, we hope that a subset of *ecoPrimers* results must correspond to them. For this purpose three different training data sets and their associated parameters were used.

The *eubact* data set contains 905 whole eubacteria genomes extracted from Genome Review release 115 (http://www.ebi.ac.uk/GenomeReviews) (31). They correspond to 603 species belonging to 311 genera. Their median size is 3.5 Mb. To identify barcodes similar to those used in bacterial biodiversity studies of soil (33), *ecoPrimers* was run on this data set using default parameters and searching for a marker of size smaller than 1 Kb (kilobases). The *e* parameter was set to 3.

The *chloro* data set contains 175 whole chloroplast genomes extracted from Genbank using eutils web api (http://eutils.ncbi.nlm.nih.gov) in January 2010. They correspond to 174 species belonging to 145 genera. From these sequences 119 belong to Tracheophyta (vascular plants, NCBI Taxid: 58023) corresponding to 118 species in 93 genera. The median size of the 175 sequences is 152 Kb. In order to find markers useful for environmental studies on vascular plant biodiversity (15), *ecoPrimers* was run on this data set with the default parameters, searching for markers with a size ranging from 10 bp to 120 bp. The *e* parameter was set to 3. The search was taxonomically restricted to Tracheophyta.

The *mito* data set is composed of 2044 whole mitochondrion genomes extracted from Genbank using eutils web api. They correspond to 2002 species belonging to 1549 genera. Among these sequences 1293 belong to Vertebrata (NCBI Taxid: 7742) corresponding to 1261 species in 966 genera. The median size of the 2044 sequences is 16.6 Kb. To search for markers usable in diet analysis studies of Carnivora, *ecoPrimers* was run on this data set with the default parameters, looking for markers with a size ranging from 50 bp to 120 bp. The *e* parameter was set to 3. On this data set two taxonomical restrictions were used. The first restricts the example sequence set E_S to NCBI Taxid: 7742 (Vertebrata) to optimize primers for vertebrates. The second defines the C_S counterexample sequence set to NCBI Taxid: 1 (Root) requiring that primers not match on sequences belonging to non-vertebrates.

In silico primer checking

Primers were checked against full Nucleic EMBL Standard release 103 database using the electronic PCR software *ecoPCR* (10). The resulting *ecoPCR* output file contains all data about potentially amplified sequences, among them the size of the amplicon, the number of mismatches associated to each primer and the taxa associated with the amplified sequences.

Empirical primer testing

Empirical testing was done for only one primer pair, named 12S-V5. This primer pair was designed by *ecoPrimers* when run on the *mito* data set with the above mentioned parameters. This primer pair had reasonably high values of B_c and B_s indices with relatively short amplification length as shown in Table 3, making it suitable for amplification from degraded DNA. 12S-V5 primer pair was empirically tested in diet analysis of three felid species, namely snow leopard (*Uncia uncia*), common leopard (*Panthera pardus*) and leopard cat (*Prionailurus bengalensis*) using feces as a source of DNA. The feces sampling was done by field workers of The Snow Leopard Trust (http://www.snowleopard.org). Snow leopard feces were collected from Mongolia in 2009 while common leopard and leopard cat feces were collected from Pakistan in 2008.

DNA extractions were performed from about 15 mg of feces with the DNeasy Blood and Tissue Kit (QIAgen GmbH, Hilden, Germany) and recovered in a total volume of 250 µl. Amplifications were carried out in a final volume of 25 µl, using 2 µl of DNA extract as template. The amplification mixture contained 1 U AmpliTaq® Gold DNA Polymerase (Applied Biosystems, Foster City, CA, USA), 10 mM Tris–HCl, 50 mM KCl, 2 mM $MgCl_2$, 0.2 mM of each dNTP, 0.1 µM of each primer (12SV05F/R), and 5 µg bovine serum albumin (BSA, Roche Diagnostic, Basel, Switzerland). The PCR mixture was denatured at 95°C for 10 min, followed by 45 cycles of 30 s at 95°C, and 30 s at 60°C; as the target sequences are shorter than 120 bp, the elongation step was removed to reduce the +A artifact (34,35) that might decrease the efficiency of the first step of the sequencing process (blunt-end ligation). The sequencing was carried out on an Illumina/Solexa Genome Analyzer IIx (Illumina Inc., San Diego, CA 92121, USA), using the Paired-End Cluster Generation Kit V4 and the Sequencing Kit V4 (Illumina Inc., San Diego, CA 92121, USA), and following manufacturer's instructions. A total of 108 nucleotides were sequenced on each extremity of the DNA fragments.

The sequence reads were analyzed using the OBITools software (http://www.prabi.grenoble.fr/trac/OBITools). First, the direct and reverse reads corresponding to a single molecule were aligned and merged using the solexaPairEnd program, taking into account data quality during the alignment and the consensus computation. Then, primers and DNA tag identifying samples were identified using the ngsfilter program. The amplified regions, excluding primers, were kept for further analysis. Strictly identical sequences were clustered together using the obiuniq program. Sequences shorter than 10 bp, or containing degenerated IUPAC nucleotide codes (other than A, C, G and T), or with occurrence less than or equal to 10 were excluded using the obigrep program. Taxon assignment was achieved using the ecoTag program (9). EcoTag relies on a dynamic programming global alignment algorithm (32) to find highly similar sequences in the reference database. This database was built by extracting the region between the two primers 12S-V5 of the mitochondrial 12S gene from EMBL nucleotide library using the output of the ecoPCR program, allowing a maximum of three mismatches between each primer and its target (10).

All computations were done on a LINUX DELL server with 32 GB of RAM (Random Access Memory).

RESULTS

Empirical testing of *ecoPrimers* on a large data set

The ability of *ecoPrimers* to analyze full genome data sets, allowing it to identify barcodes without *a priori* targeting of any potential locus, relies on its algorithm efficiency. Efforts have been made during algorithm conception both in terms of memory and time. We have empirically estimated the memory requirements of SPA and compared it with three algorithms *KMR* (23), Suffix trees (24) and Suffix arrays (25). Memory and time complexities were estimated using *eubact* as data set. Size of L_P list and computation time was measured after each sequence insertion during SPA execution.

SPA without data mining. The program was first run without data mining. Figure 2a displays the evolution of L_P size. As expected, it increased during the insertion of the first 273 sequences. The limit value corresponds to $|E_s| - Q_s + 1$. At this point, many words could not reach Q_s and were discarded from L_P. The maximum size of L_P is about 7.8 GB for 3 Gb of sequences. This corresponds to a usage of about 3.6 bytes per nucleotide analyzed on both strands, including one byte to store the sequence itself. This is already better than the three standard algorithms, but this transient long list has a drastic impact on memory and speed performances. Time evolution during execution (Figure 2c) evolves in a quadratic way with the sequence count. Theoretically, in the worst case, the algorithm has a complexity of $O(N^2)$ during this phase, where N is count of processed sequences. Then time evolves linearly, as $|L_P|$ becomes very small. With *eubact* data set, total time used for the strict primer algorithm is about 1 h and 40 min.

SPA with data mining. The experiment was repeated with data mining activated. This time the majority of hashed words were not included in the L_W list because they occurred in less than Q_s sequences of E_s. The effect of this reduction of $|L_W|$ is observable on Figure 2b. The memory size of L_P is never over 2.5 KB (less than 210 patterns). The global size used with data mining including H_s, L_P, L_W and the sequence itself is about 1.1 bytes per nucleotide. The second effect of this drastic size reduction of L_P and L_W is the speed increase. With data mining the execution time of the strict primer detection is about 5 min (2 min for H_m building and 3 min for strict primer detection). Moreover empirical time complexity is now linear with the count of sequences (Figure 2d).

Global execution. A full search for primers using data mining on the *eubact* data set is about 3 h 40 min. Main time is devoted to the agrep algorithm. Execution time of this part of our global algorithm is in $O((|E_s + |C_s|)|L_P|)$. On this data set *ecoPrimers* never used more than 4 GB of memory.

Designed primers. A Eubacteria training data set was used to demonstrate efficiency of the algorithm, so primers identified with this data set were not checked further. The program proposed almost 5521 primer pairs. Out of these 5521 primer pairs, we investigated the first few pairs and they seem to amplify part of functional RNA genes (rRNA 16S gene, rRNA 23S genes). The five pairs are presented in Table 1, they all correspond to parts of the 16S gene.

Validation of *ecoPrimers* on vascular plants

As the majority of already published barcodes for plants correspond to regions of the chloroplast DNA (4,15,16), we ran *ecoPrimers* on the *chloro* data set. Three hundred and forty three primer pairs were selected out of 265 273 primer pairs identified limiting the value of *barcode specificity* to at least 50%. The specified parameters allow the selection of markers with properties similar to that of g/h primers (15). These primers have already been used for

Table 1. The five best primer pairs proposed by *ecoPrimers* to amplify potential barcode markers specific of eubacteria

Sequences		T_m		Amplified E_s	B_c	B_s	Fragment size (bp)			Region
Direct	Reverse	P1	P2				Min	Max	Average	
CGACACGAGCTGACGACA	CTACGGGAGGCAGCAGTG	60.5	60.8	603	1.00	0.927	668	987	699.07	16S RNA
CTACGGGAGGCAGCAGTG	GGTATCTAATCCTGTTTG	60.8	47.5	603	1.00	0.910	392	708	417.52	16S RNA
CTACGGGAGGCAGCAGTG	GCGGGCCCCCGTCAATTC	60.8	64.9	603	1.00	0.907	525	844	556.49	16S RNA
AGCAGCCGCGGTAATACG	GCGGGCCCCCGTCAATTC	61.1	64.9	603	1.00	0.842	370	666	380.21	16S RNA
ACCGCGGCTGCTGGCACG	CTACGGGAGGCAGCAGTG	69.6	60.8	603	1.00	0.819	128	598	152.66	16S RNA

Amplified E_s column indicates electronically amplified species count belonging to the Eubacteria data set.

Table 2. The five best primer pairs proposed by *ecoPrimers* to amplify potential barcode markers specific of vascular plants

Primer name	Sequences		T_m		Amplified E_s	B_c	B_s	Fragment size (bp)			Region
	Direct	Reverse	P1	P2				Min	Max	Average	
similar to *g/h*	GGCAATCCTGAGCCAAAT	TGAGTCTCTGCACCTATC	56.1	53.5	114	0.966	0.711	10	90	45.65	*trn*L-P6-loop
similar to *g/h*	ATTGAGTCTCTGCACCTA	GGGCAATCCTGAGCCAAA	52.7	58.4	114	0.966	0.658	13	93	48.65	*trn*L-P6-loop
similar to *g/h*	AGCTTCCATTGAGTCTCT	GGGCAATCCTGAGCCAAA	53.0	58.4	111	0.941	0.649	20	100	55.96	*trn*L-P6-loop
	TGGTTATTTACTAAAATC	TTTGGTTAAGATATGCCA	41.9	48.9	116	0.983	0.647	100	103	100.3	*psb*CL
	GCAATCCTGAGCCAAATC	GCTTCCATTGAGTCTCTG	54.8	53.4	112	0.949	0.652	17	97	52.73	*trn*L

g/h primers were proposed by Taberlet *et al.* (15) for vascular plant identification. Amplified E_s column indicates electronically amplified species count belonging to the vascular plant example set.

Table 3. The five best primer pairs proposed by *ecoPrimers* to amplify potential barcode markers specific of vertebrates

Primer Name	Sequences		T_m		Amplified		B_c	B_s	Fragment size (bp)			Region
	Direct	Reverse	P1	P2	E_s	C_s			Min	Max	Average	
	ACTGGGATTAGATACCCC	TAGAACAGGCTCCTCTAG	52.6	52.3	1221	31	0.968	0.858	85	117	105.38	16S RNA
12S–V5	TAGAACAGGCTCCTCTAG	TTAGATACCCCACTATGC	52.3	50.7	1236	7	0.980	0.720	73	110	98.32	12S RNA
	AGGGATAACAGCGCAATC	TCGTTGAACAAACGAACC	55.6	54.4	1256	18	0.996	0.459	63	84	82.03	12S RNA
similar to 16Sr	CTCCGGTCTGAACTCAGA	GATGTTGGATCAGGACAT	56.1	52.1	1253	59	0.994	0.196	53	59	58.22	16S RNA
	ATGTTGGATCAGGACATC	CTCCGGTCTGAACTCAGA	52.1	56.1	1253	35	0.994	0.195	54	60	57.22	16S RNA

16Sr primers were proposed by Palumbi *et al.* (14) for mammal identification (37). Amplified E_s and C_s columns indicate electronically amplified species counts belonging respectively to the vertebrate example set and to the non-vertebrate counterexample set.

several metabarcoding applications, such as diet analysis (9,36) or to reconstruct past arctic vegetation (6). Table 2 presents the five primers pairs selected from five best regions identified by *ecoPrimers*. Not only did *ecoPrimers* identify primers similar to g/h as expected, amplifying the same *trn*L P6-loop, but it ranked them with the best mark. Most of the primer pairs amplify regions of functional RNA genes, or of introns. (34 primers amplify regions of *trn*L, 41 primers amplify regions of *trn*W, 11 primers amplify regions of *trn*Y and 13 primer amplify regions of *trn*H. Finally 231 primer pairs amplify regions of protein coding genes including *psa*B, *psa*A, *psb*A, *psb*C and the intergenic region of *psb*L and *psb*F).

Validation of *ecoPrimers* on vertebrates

In a similar way as we did for vascular plants, we ran *ecoPrimers* on the *mito* data set, asking for primers amplifying only Vertebrata.

Designed primers. Forty-two primer pairs were identified. As for previous tests, they were mainly located on non-protein coding sequences (30 in rRNA 16S gene, 12 in rRNA 12S gene). The five best primer pairs are presented in Table 3. The first of them, named *12S-V5*, was more carefully checked using bioinformatics and experimental approaches (see below). The third and fourth correspond to variants of primers amplifying a region of the 16S rRNA gene already proposed as barcode marker for mammals (14,37)

Bioinformatics validation of the 12S-V5 primer pair. The *12S-V5* primer pair amplifies a part of the 12S rRNA gene including its V5 variable region. The amplified region from the *ecoPrimers* results range from 73 bp to 110 bp. It is able to amplify 98% of the sequence training set (B_c = 0.98) and unambiguously identifies 74% of those amplified species (B_s = 0.74). Only 7 taxa of over 741 represented in the counterexample set of sequences C_S are recognized by this primer pair. Better estimation of the quality of this barcode was achieved using *ecoPCR* against EMBL nucleotide database (10). We set *ecoPCR* parameters to allow *in silico* PCR amplification ranging from a size between 50 bp to 250 bp with no more than 3 mismatches per primer. It resulted in the potential amplification of 17737 sequences of vertebrate (according to the EMBL annotation) and only 79 sequences belonging to other taxa. Of these non-vertebrate sequences, 66 of them belong to the Crustacea (NCBI Taxid: 6657), 5 belong to Insecta (NCBI Taxid: 50557), 3 belong to Arthropoda (NCBI Taxid: 6656) and 1 sequence belongs to each of the following taxa: Gastropoda (NCBI Taxid: 6448), Lineidae (NCBI Taxid: 6222), Loxosomatidae (NCBI Taxid: 231594). All these non-vertebrate taxa present two or three mismatches with both primers. The two last non-vertebrate sequences exhibit zero or one mismatch for both primers but they correspond to mis-assigned taxa. The first one embl:EU626452, annotated as an uncultured bacterium (NCBI Taxid: 77133), is identical to a human sequence. The second one embl:AF257243, annotated as a nematode (*Onchocerca volvulus* NCBI Taxid: 6282), is similar to many bony fish (Actinopterygii NCBI Taxid: 7898) sequences. The amplified vertebrate sequences correspond to 5926 species and 2732 genera. Among them 4537 species ($B_s = 0.77$) and 2430 genera ($B_s = 0.89$) are unambiguously identified. Among the 17737 sequences of vertebrate only 353 have two or three mismatches with the both primers. A total of 266 of them belong to reptiles (Sauropsida NCBI Taxid: 8457), 24 sequences belong to amphibians (Amphibia NCBI Taxid: 8292) and 3 sequences belong to the Batrachoididae family (NCBI Taxid: 8065). The 60 remaining sequences belong to mammals (NCBI Taxid: 40674) but most of these sequences are annotated as a nuclear copy of this mitochondrial locus. Table 4 resumes the distribution of mismatches of the two 12S-V5 primers among vertebrate species.

Table 4. Number of vertebrate species exhibiting from 0 to 3 mismatches for forward and reverse 12S-V5 primers

Number of mismatches	Number of species	
	Forward primer	Reverse primer
0	3272	4592
1	2031	1021
2	465	291
3	158	20

Experimental validation of primer 12S-V5. The empirical testing of the 12S-V5 primer pair was carried on felid feces, to assess their diet. One, one and two feces were used for snow leopard (*U. uncia*), common leopard (*P. pardus*) and leopard cat (*P. bengalensis*), respectively. The results are summarized in Table 5. As expected, both felid (i.e. predator) and the prey sequences were obtained. The B_s of the amplified sequences allowed us to unambiguously distinguish the three predators, and to identify different prey, including three mammals, one bird and one amphibian.

DISCUSSION

In this article, we have clearly demonstrated the ability of the *ecoPrimers* software to fulfill all the requirements for designing new barcode regions suitable for metabarcoding studies. This software has the ability to scan large training databases (example and counterexample sets) so as to design highly conserved primers that have the potential to amplify a variable DNA region. The ranking of the primer pairs is based on the two previously proposed indices B_c and B_s (10) that evaluate the taxonomic range potentially amplified by a primer pair, and the discrimination capacity of the amplified region, respectively. A large set of parameters can be specified for tuning the algorithm, including (i) the maximum number of errors allowed between each primer and the target sequence, (ii) the possibility to restrict the search to a given taxonomic level (example set), (iii) the possibility to define a set of counterexample taxa that the primers should not amplify (within or outside of the clade used for the search), (iv) the minimum and maximum length of the amplified region, (v) the possibility to consider that the database sequences are circular, (vi) the possibility to

Table 5. Count of sequences observed per sample after Solexa sequencing of 4 PCR amplicons

		Feces			
		Common leopard	Snow leopard	Leopard cat	
				1	2
Predator	Common leopard (*P. pardus*)	2460	–	–	–
	Snow leopard (*U. uncia*)	–	10 807	-	-
	Leopard cat (*P. bengalensis*)	–	–	1982	9765
Prey	Domestic goat (*Capra hircus*)	2969	–	–	–
	Siberian ibex (*Capra sibirica*)	–	1256	–	–
	Shrew (*Crocidura pullata*)	–	–	–	964
	Chukar partridge (*Alectoris chukar*)	–	–	1711	
	Muree hill frog (*Paa vicina*)	–	–	–	982

Each of them corresponds to one predator feces.

require a strict match on a specified number of nucleotides on 3′-end of the primers, (vii) the proportion of strict matching primers on the example set, (viii) the proportion of primers matching with specified number of errors on the example set, (ix) the proportion of primers matching the counterexample dataset, and finally (x) the possibility of avoiding primers matching more than once in one sequence of the example set. The efficiency of *ecoPrimers* has been successfully validated, both via bioinformatics analyses and via empirical experiments.

The main advantage and the originality of *ecoPrimers* is its full integration of the taxonomy. This characteristic has been implemented in a way that allows the design of new barcodes specific to any taxonomic group, as well as the optional exclusion of any other clades. For example, if analyzing the fish diet of an otter (genus *Lutra*) using their feces, it is possible with *ecoPrimers* to design a short barcode that includes all teleost fish (Teleostei) and excludes the genus *Lutra*; such a strategy will not only promote prey DNA amplification, but also prevent otter DNA amplification. Another key advantage is the speed efficiency of the *ecoPrimers* algorithm when it is used on whole mitochondrial or chloroplast genomes as example sets, and its ability to run on other huge data sets like whole eubacteria genomes.

ecoPrimers is particularly useful for setting up the analysis of environmental samples using a metabarcoding approach. In such a situation, to avoid amplification bias among the different taxonomic groups, it is extremely important to work with highly conserved primers. Unfortunately, for higher taxonomic group (e.g. vertebrate, angiosperms) it is impossible to find primer pairs amplifying all species without mismatch (B_c) and with a good specificity (B_s). So we cannot exclude that some species could be missed by a primer pair. To limit potential problems related to relatively low coverage of a primer pair, it could be useful to analyze the same sample with several markers targeting the same taxonomic group.

The possibility to choose the length of the barcode is crucial when working with degraded DNA: in such a context only fragments shorter than 100 bp can be reliably amplified. According to our experience, in some taxonomic groups, it is even possible to design extremely short barcodes that nevertheless have a very high coverage and specificity. This is the case for earthworms (Lumbricina) where a 30 bp barcode located on the mitochondrial 16S gene allows the identification of all species from the French Alps analyzed up to now (Bienert *et al.*, submitted for publication). Even when using good quality DNA, the length of the sequence reads obtained from the DNA sequencer might impose a maximum length when designing new barcodes. The current standardized barcodes for animals (38) and plants (4) were designed according to the technological characteristics of the sanger sequencing using capillary electrophoresis (sequence reads shorter than 1 kb). In the near future, if the read length of next generation DNA sequencers increases to several kilobases, it might be worthwhile to redesign much longer barcodes to significantly increase the taxonomic resolution. As more and more whole mitochondrial and chloroplast genomes become available, *ecoPrimers* has the potential to provide new optimal barcodes.

The majority of barcodes proposed by *ecoPrimers* for Eubacteria, vascular plants and vertebrates are located on ribosomal DNA. The only exception was on chloroplast DNA, with a few primers located either on transfer RNA or on protein genes. As a consequence, the example set of sequences can be taxonomically enlarged by only taking into account the ribosomal genes, and not the whole mitochondrial or chloroplast genomes. In the same way, if the goal is to design a nuclear barcode, the nuclear ribosomal genes can be efficiently used as the example set.

According to our experience, it is sometimes difficult to find suitable short barcodes for some taxonomic groups, particularly if they diverged a very long time ago. Usually, the higher the taxonomic level considered, the greater the difficulty to find universal barcodes. If such a problem occurs, we advise first modifying the parameters by relaxing as much as possible the different constraints, and then trying to design several barcodes, one for each of the taxonomic groups at a lower level. The other option is to degenerate the proposed primers to enlarge their taxonomic coverage. Combined use of *ecoPrimers* and *ecoPCR* (10) is convenient for this purpose.

As more and more sequences become available in public databases, by using larger example sets, *ecoPrimers* will be more and more efficient for designing new barcodes that can be precisely optimized according to the biological question and to the experimental constraints. The biological question might impose a particular level of specificity (e.g. species level), or conversely a broad taxonomic range, but with a resolution at the family level. The experimental constraints might concern the length of the barcode, or the avoidance of amplifying another non-target taxonomic group. The analysis of environmental samples using next generation sequencers is already frequently used for estimating the diversity of bacteria, e.g. (33), fungi, e.g. (39), and more recently of nematodes, e.g. (40). There are more and more research projects extending the approach to other taxonomic groups. In such a context, the availability of a program allowing the design of the most suitable barcode will probably enhance studies analyzing the biodiversity of environmental samples. *ecoPrimers* is available as an open source software at: http://www.grenoble.prabi.fr/trac/ecoPrimers.

FUNDING

Alocad project (ANR-06-PNRA-004- 02, in part): European Project EcoChange (FP6-036866, in part). The authors T.R. and W.S. were funded by HEC (Higher Education Commission), Government of Pakistan. Funding for open access charge: European Project EcoChange (FP6-036866).

Conflict of interest statement. T.R., P.T. and E.C. are co-inventors of a pending French patent on the primer pair named $12S-V5_F$ and $12S-V5_R$ and on the use of the amplified fragment for identifying vertebrate species from environmental samples. This patent only restricts commercial applications and has no impact on the use of this method by academic researchers.

REFERENCES

1. Hebert,P.D.N., Cywinska,A., Ball,S.L. and deWaard,J.R. (2003) Biological identifications through DNA barcodes. *Proc. Biol. Sci.*, **270**, 313–321.
2. Valentini,A., Pompanon,F. and Taberlet,P. (2009) DNA barcoding for ecologists. *Trends Ecol. Evol.*, **24**, 110–117.
3. Ahrens,D., Monaghan,M.T. and Vogler,A.P. (2007) DNA-based taxonomy for associating adults and larvae in multi-species assemblages of chafers (Coleoptera: Scarabaeidae). *Mol. Phylogenet Evol.*, **44**, 436–449.
4. CBOL Plant Working Group (2009) A DNA barcode for land plants. *Proc. Natl Acad. Sci. USA*, **106**, 12794–12797.
5. Willerslev,E., Hansen,A.J., Binladen,J., Brand,T.B., Gilbert,M.T.P., Shapiro,B., Bunce,M., Wiuf,C., Gilichinsky,D.A. and Cooper,A. (2003) Diverse plant and animal genetic records from Holocene and Pleistocene sediments. *Science*, **300**, 791–795.
6. Sønstebø,J.H., Gielly,L., Brysting,A.K., Elven,R., Edwards,M., Haile,J., Willerslev,E., Coissac,E., Rioux,D., Sannier,J. *et al.* (2010) Using next-generation sequencing for molecular reconstruction of past Arctic vegetation and climate. *Mol. Ecol. Resour.*, **10**, 1009–1018.
7. Ficetola,G.F., Miaud,C., Pompanon,F. and Taberlet,P. (2008) Species detection using environmental DNA from water samples. *Biol Lett.*, **4**, 423–425.
8. Valentini,A., Miquel,C., Nawaz,M.A., Bellemain,E., Coissac,E., Pompanon,F., Gielly,L., Cruaud,C., Nascetti,G., Wincker,P. *et al.* (2009) New perspectives in diet analysis based on DNA barcoding and parallel pyrosequencing: the *trn*L approach. *Mol. Ecol. Resour.*, **9**, 51–60.
9. Pegard,A., Miquel,C., Valentini,A., Coissac,E., Bouvier,F., François,D., Taberlet,P., Engel,E. and Pompanon,F. (2009) Universal DNA-based methods for assessing the diet of grazing livestock and wildlife from feces. *J. Agric Food Chem.*, **57**, 5700–5706.
10. Ficetola,G.F., Coissac,E., Zundel,S., Riaz,T., Shehzad,W., Bessiere,J., Taberlet,P. and Pompanon,F. (2010) An *In silico* approach for the evaluation of DNA barcodes. *BMC Genom.*, **11**, 434.
11. Bellemain,E., Carlsen,T., Brochmann,C., Coissac,E., Taberlet,P. and Kauserud,H. (2010) ITS as an environmental DNA barcode for fungi: an *in silico* approach reveals potential PCR biases. *BMC Microbiol.*, **10**, 189.
12. Meusnier,I., Singer,G.A.C., Landry,J.F., Hickey,D.A., Hebert,P.D.N. and Hajibabaei,M. (2008) A universal DNA mini-barcode for biodiversity analysis. *BMC Genom.*, **9**, 214.
13. Kocher,T.D., Thomas,W.K., Meyer,A., Edwards,S.V., Pääbo,S., Villablanca,F.X. and Wilson,A.C. (1989) Dynamics of mitochondrial DNA evolution in animals: amplification and sequencing with conserved primers. *Proc. Natl Acad. Sci. USA*, **86**, 6196–6200.
14. Palumbi,S. (1996) Nucleic acids II: the polymerase chain reaction. In: Hillis,D., Moritz,C. and Mable,B. (eds), *Molecular Systematics*, 2nd edn. Sinauer Assoc., Sunderland, MA, pp. 205–247.
15. Taberlet,P., Coissac,E., Pompanon,F., Gielly,L., Miquel,C., Valentini,A., Vermat,T., Corthier,G., Brochmann,C. and Willerslev,E. (2007) Power and limitations of the chloroplast *trn*L (UAA) intron for plant DNA barcoding. *Nucleic Acids Res.*, **35**, e14.
16. Kress,W.J., Wurdack,K.J., Zimmer,E.A., Weigt,L.A. and Janzen,D.H. (2005) Use of DNA barcodes to identify flowering plants. *Proc. Natl Acad. Sci. USA*, **102**, 8369–8374.
17. Rozen,S. and Skaletsky,H. (2000) Primer3 on the WWW for general users and for biologist programmers. *Methods Mol. Biol.*, **132**, 365–386.
18. Kim,N. and Lee,C. (2007) QPRIMER. *Bioinformatics*, **23**, 2331–2333.
19. Bode,M., Khor,S., Ye,H., Li,M.-H. and Ying,J.Y. (2009) TmPrime: fast, flexible oligonucleotide design software for gene synthesis. *Nucleic Acids Res.*, **37**, W214–W221.
20. Bekaert,M. and Teeling,E.C. (2008) UniPrime: a workflow-based platform for improved universal primer design. *Nucleic Acids Res.*, **36**, e56.
21. Duitama,J., Kumar,D.M., Hemphill,E., Khan,M., Mandoiu,I.I. and Nelson,C.E. (2009) PrimerHunter: a primer design tool for PCR-based virus subtype identification. *Nucleic Acids Res.*, **37**, 2483–2492.
22. Jarman,S.N. (2004) Amplicon: software for designing pcr primers on aligned dna sequences. *Bioinformatics*, **20**, 1644–1645.
23. Karp,R.M., Miller,R.E. and Rosenberg,A.L. (1972) *STOC '72: Proceedings of the fourth annual ACM symposium on Theory of computing*. ACM, New York, NY, USA, pp. 125–136.
24. McCreight,E.M. (1976) A space-economical suffix tree construction algorithm. *J. ACM*, **23**, 262–272.
25. Manber,U. and Myers,G. (1990) *SODA '90: Proceedings of the first annual ACM-SIAM symposium on Discrete algorithms*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, pp. 319–327.
26. Kurtz,S. and Schleiermacher,C. (1999) REPuter: fast computation of maximal repeats in complete genomes. *Bioinformatics*, **15**, 426–427.
27. Achaz,G., Boyer,F., Rocha,E.P.C., Viari,A. and Coissac,E. (2007) Repseek, a tool to retrieve approximate repeats from large DNA sequences. *Bioinformatics*, **23**, 119–121.
28. Dijkstra,E.W. (1982) Smoothsort, an alternative for sorting in situ. *Sci. Comput. Program*, **1**, 223–233.
29. Wu,S. and Manber,U. (1992) Agrep, a fast approximate pattern-matching tool. In *Proceedings USENIX Winter 1992 Technical Conference*, pp. 153–162.
30. Santalucia,J. and Hicks,D. (2004) The thermodynamics of DNA structural motifs. *Annu. Rev. Biophys. BioMol. Struct*, **33**, 415–440.
31. Kersey,P., Bower,L., Morris,L., Horne,A., Petryszak,R., Kanz,C., Kanapin,A., Das,U., Michoud,K., Phan,I. *et al.* (2005) Integr8 and Genome Reviews: integrated views of complete genomes and proteomes. *Nucleic Acids Res.*, **33**, D297–D302.
32. Needleman,S.B. and Wunsch,C.D. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.*, **48**, 443–53.
33. Edwards,R.A., Rodriguez-Brito,B., Wegley,L., Haynes,M., Breitbart,M., Peterson,D.M., Saar,M.O., Alexander,S., Alexander,E.C. Jr and Rohwer,F. (2006) Using pyrosequencing to shed light on deep mine microbial ecology. *BMC Genom.*, **7**, 57.
34. Brownstein,M.J., Carpten,J.D. and Smith,J.R. (1996) Modulation of non-templated nucleotide addition by Taq DNA polymerase: primer modifications that facilitate genotyping. *Biotechniques*, **20**, 1004–1006, 1008–1010.
35. Magnuson,V.L., Ally,D.S., Nylund,S.J., Karanjawala,Z.E., Rayman,J.B., Knapp,J.I., Lowe,A.L., Ghosh,S. and Collins,F.S. (1996) Substrate nucleotide-determined non-templated addition of adenine by Taq DNA polymerase: implications for PCR-based genotyping and cloning. *Biotechniques*, **21**, 700–709.
36. Soininen,E.M., Valentini,A., Coissac,E., Miquel,C., Gielly,L., Brochmann,C., Brysting,A.K., Sonstebo,J.H., Ims,R.A., Yoccoz,N.G. *et al.* (2009) Analysing diet of small herbivores: the efficiency of DNA barcoding coupled with high-throughput pyrosequencing for deciphering the composition of complex plant mixtures. *Front Zool.*, **6**, 16.
37. Palumbi,S., Martin,A., Romano,S., McMillan,W., Stice,L. and Grabowski,G. (1991) *The Simple Fool's Guide to PCR, Version 2.0*. University of Hawaii, Honolulu.

38. Hebert,P.D.N., Ratnasingham,S. and deWaard,J.R. (2003) Barcoding animal life: cytochrome c oxidase subunit 1 divergences among closely related species. *Proc. Biol. Sci.*, **270**, S96–S99.
39. Opik,M., Metsis,M., Daniell,T.J., Zobel,M. and Moora,M. (2009) Large-scale parallel 454 sequencing reveals host ecological group specificity of arbuscular mycorrhizal fungi in a boreonemoral forest. *New Phytol.*, **184**, 424–437.
40. Porazinska,D.L., Giblin-Davis,R.M., Faller,L., Farmerie,W., Kanzaki,N., Morris,K., Powers,T.O., Tucker,A.E., Sung,W. and Thomas,W.K. (2009) Evaluating high-throughput sequencing as a method for metagenomic analysis of nematode diversity. *Mol. Ecol. Resour.*, **9**, 1439–1450.

3.1 Dealing With PCR Errors In The Computation Of B_s

As discussed in chapter 2 section 2.3.4 due to PCR errors we may not clearly distinguish two taxa exhibiting only one or two differences between their barcodes. To more accurately estimate B_s in a way considering this possibility, we can put a certain threshold on the distance that two sequences need to exhibit in order to be declared as distinguishable. With this approach those erroneous sequences which were well identified before will no longer be unambiguously identified, thus lowering the falsely increased value of B_s. The algorithm can be used with option $-e$ in $ecoTaxSpecifity$ program present in OBITools.[1] Table 3.1 shows the change in barcode specificity B_s for some already published primer pairs if all the sequences at a distance $d = 1$ or $d = 2$ are considered similar.

Primer Name	Sequences		Amplified E_s	Well Identified E_s	d	B_s
	Direct	Reverse				
$12S - V5$	TAGAACAGGCTCCTCTAG	TTAGATACCCCACTATGC	1182	1006	0	0.85
$12S - V5$	TAGAACAGGCTCCTCTAG	TTAGATACCCCACTATGC	1182	884	1	0.74
$12S - V5$	TAGAACAGGCTCCTCTAG	TTAGATACCCCACTATGC	1182	773	2	0.65
similar to $16Sr$	CTCCGGTCTGAACTCAGA	GATGTTGGATCAGGACAT	1253	243	0	0.19
similar to $16Sr$	CTCCGGTCTGAACTCAGA	GATGTTGGATCAGGACAT	1253	90	1	0.07
similar to $16Sr$	CTCCGGTCTGAACTCAGA	GATGTTGGATCAGGACAT	1253	40	2	0.03
similar to g/h	AGCTTCCATTGAGTCTCT	GGGCAATCCTGAGCCAAA	111	78	0	0.70
similar to g/h	AGCTTCCATTGAGTCTCT	GGGCAATCCTGAGCCAAA	111	61	1	0.55
similar to g/h	AGCTTCCATTGAGTCTCT	GGGCAATCCTGAGCCAAA	111	58	2	0.52

Table 3.1: The value of B_s is shown for some standard and newly published primer pairs with d equal to 0, 1 and 2. *mito* and *chloro* datasets were used restricting the search to Vertebrata (NCBI Taxid: 7742) and Tracheophyta (vascular plants, NCBI Taxid: 58023).

3.2 One Step Ahead In Metabarcoding: The Sets Approach

For metabarcoding applications, those involving environmental samples and ancient DNA, it is extremely difficult to amplify regions longer than 150 bp. Mostly barcode markers amplifying regions up to 100 bp or shorter are preferred. However for such a short amplification length, the level of inter and intra species variation may not be enough since the resolution capacity is directly dependent on amplification length. In order to well identify a large part of the individuals present in a given environmental sample, one idea could be to use a set of barcode markers instead of a single marker. This set approach can also be interesting for combining even short barcode markers like those between $10 - 60$ bp of amplification length. Moreover the sets can also be useful for the taxa where universal barcode design is difficult like those for *insecta*. Such a set could be designed by choosing the best barcode regions and primer pairs such that the coverage and specificity of whole set is maximized. In order to decide that which set is the best, we may need to

[1] https://www.grenoble.prabi.fr/OBITools

make all possible number of solution sets from our solution space and compare them. The solution space, in this case, is the big set of all barcode markers available, for example all barcode markers designed by *ecoPrimers* for our *mito* data set. Any subset of this set represents a potential solution set.

3.2.1 Problem Statement And Complexity:

If U_P is the set of all primer pairs (or equivalently barcode markers) then find the smallest set $S_P \subset U_P$ that maximizes B_s and B_c. Maximization of both B_s and B_c falls in the category of set cover problem which is a well known NP-complete problem and one of Karp's 21 NP-complete problems (Karp, 1972). NP-completeness implies that finding an exact solution simply requires to evaluate all possible subsets of U_P, where the number of these sets is $2^{|U_P|}$ (number of elements of power set of U_P). Hence finding an exact solution for even a moderate sized U_P is infeasible.

Although finding an exact solution is infeasible, however we can develop techniques using some existing metaheuristics (*e.g.* Simulated Annealing and Tabu Search) to find potential good sets. In order to use these heuristics, it is required to define quantitative ways to compare two solutions to decide which one is better. Usually some energy function is used for this purpose and a set having minimum energy is better than the others. Since in our case both of the factors involved need be maximized for target good set, so we can equivalently define some score function that will assess the goodness of the set. A set that maximizes the score function will be better than others.

3.2.2 Score Function

We have proposed a very simple score function that is based on the maximization of coverage.

$$Set_{score} = B_c$$

Our strategy is as follows: Design the primer pairs which amplify only a few taxa (*i.e.* associated to a low B_c value) but with a high B_s value (*i.e.* close to 1.0). And then use these low coverage and highly specific primer pairs as solutions space for metaheurisitc to find out the optimal solution set. The reason for such an approach is that for certain taxa like *metazoas* and *insecta*, universal primer design is very difficult and mostly highly degenerated primer pairs have been proposed. This is because these taxa does not share many highly conserved regions. Thus for such taxa if low coverage primers are designed, the union of coverage can be maximized by combining several primer pairs.

3.2.3 Design Of Low Coverage Primers

As said earlier, primer design is the problem of finding highly conserved regions (repeats) among a set of sequences. In our *ecoPrimers* algorithm, we first look for strict repeats and then use those strict repeats as pattern in *Agrep* algorithm to find out approximate versions of strict repeats. Such a scheme is used because primer sequence and matrix sequences may allow some errors. The affect of this tolerance of errors is the increase in the coverage of a primer pair. In our sets approach we have decided to build primers which have a very low value of B_c in order to maximize the union of set. However since we want to maximize B_c using sets approach, we actually no longer need to use the *Agrep* algorithm. And we can use all the low coverage strict repeats for pairing and further processing. But one problem in this approach is that, during our strict repeats finding algorithm, as explained in *ecoPrimers* publication, we do not save the positions of repeats because they are recomputed during *Agrep* part, and pairing cannot be performed if we do not know the positions of primer sequences on the DNA sequence. Since a large number of low coverage repeats can be found (because there are more words which are present in 5% of sequences than those which are present in 50% of the sequences), using the *Agrep* algorithm to compute the positions of primers on DNA sequences will be highly expensive in time computation. In order to find the positions of low coverage strict repeats, thus instead of using *Agrep* algorithm, we make use of automata approach by implementing an $Aho - Corasick$ algorithm (Aho and Corasick, 1975) in order to find out the positions of strict repeats. $Aho - Corasick$ is a dictionary-matching algorithm that locates elements of a finite set of strings (the "dictionary", strict repeats in our case) within an input text. The complexity of the algorithm is linear in the length of the patterns plus the length of the searched text plus the number of output matches. It matches all patterns simultaneously, thus it is very fast. Once we have the positions, we use the positions to locate the primer sequences on actual DNA sequence and to perform the primer pairing step.

3.2.4 Reducing The Search Space

Since we design primers which have very low coverage (*e.g.* primers amplifying only 5% of total taxa) as a result we get a huge number of primer pairs. This large number of primer pairs increase the search space for metaheuristics, thus increasing the running time of metaheuristic algorithm. We have developed an efficient strategy to reduce the size of search space. This strategy is based on graphs approach. We define a graph $G(P, L)$,where P is the set of all the primer pairs identified by *ecoPrimers*. If t_i is the set of taxa identified by the primer pair $p_i \in P$ and t_j is the set of taxa identified by the primer pair $p_j \in P$, then we define $l_{i,j} = t_i \cap t_j$. Thus L the relation defining the graph can be expressed as:

$$L(p_i, p_j) = \begin{cases} & \lceil \min(|p_i|, |p_j|) \times 0.05 \rceil \quad \geq |l_{l,j}| \\ \mathit{and} & \lceil \min(|p_i|, |p_j|) \times 0.5 \rceil \quad \leq |l_{l,j}| \end{cases} \tag{3.2.1}$$

So an edge between any two nodes p_i and p_j exists if the above relation is true. Each connected component from this graph G can be considered as search space for metaheuristics.

Although we have not implemented it, but there is another idea to further reduce the search space. To achieve this, we can calculate the upper bound of B_c for each component, then by selecting the component having maximum value of B_c we can find maximal cliques in this component using algorithm developed by (Born and Kerbosch, 1973) by integrating Tabu Search or Simulated Annealing for finding neighbors in this algorithm. The found maximal clique will serve as new search space for metaheuristics for finding the final solution set.

3.2.5 Neighboring Criteria For Metaheuristics

In our implementation we generate different sets of sizes 3 to 10 using Simulated Annealing and Tabu Search heuristics. These heuristics need some neighboring criteria to generate a new set from seed set. Using these criteria new set is chosen in the neighborhood of the old one. We have implemented and experimented with following four neighboring criteria:

- *All Random:* Replace random number of elements from the seed set with random elements not already in the seed set to get a new neighbor set.
- *Random with least contributing elements:* Replace a random number of least contributing elements from the seed set with the random elements from the remaining set to get new neighbor set.
- *One with next:* Replace only one element from the seed set with the next element in the remaining set to get new neighbor set.
- *Least contributing with the next:* Replace the least contributing element from the seed set with the next element in the remaining set to get new neighbor set.

Mostly All Random criterion gives better sets with higher scores. Moreover it is observed that both Simulated Annealing and Tabu Search converge to same best solution at the end.

3.2.6 Results

The results for sets approach are mainly based on *metazoa*. *ecoPrimers* was run on *mito* data set that comprises of 2044 whole mitochondrion genomes extracted from *Genbank* using eutils web api.[2]

#	Sequences		Tm		Amplified E_s	B_c	B_s	Fragment size (bp)		
	Direct	Reverse	P1	P2				min	max	average
0	ACACCGCCCGTCACTCTC	TTACCATGTTACGACTTG	62.5	50.1	563	0.260	0.840	45	59	51.96
1	CACACCGCCCGTCACTCT	TTACCATGTTACGACTTG	62.8	50.1	562	0.259	0.840	46	59	52.95
2	ACACACCGCCCGTCACTC	TTACCATGTTACGACTTG	63.1	50.1	559	0.258	0.839	47	59	53.92
3	ACACCGCCCGTCACTCTC	TACCATGTTACGACTTGC	62.5	52.9	547	0.252	0.835	44	59	50.96
4	ACACCGCCCGTCACTCTC	ACCATGTTACGACTTGCC	62.5	55.8	547	0.252	0.835	43	59	49.98
5	CACACCGCCCGTCACTCT	CCATGTTACGACTTGCCT	62.8	55.5	546	0.252	0.835	43	59	49.97
6	CACACCGCCCGTCACTCT	CATGTTACGACTTGCCTC	62.8	54.2	546	0.252	0.835	42	58	48.97
7	ACACACCGCCCGTCACTC	ACCATGTTACGACTTGCC	63.1	55.8	545	0.251	0.835	45	59	51.95
8	ACACACCGCCCGTCACTC	TACCATGTTACGACTTGC	63.1	52.9	545	0.251	0.835	46	59	52.93
9	ACACACCGCCCGTCACTC	ATGTTACGACTTGCCTCC	63.1	55.2	538	0.248	0.838	42	58	48.94
10	ACACCGCCCGTCACTCTC	CTTACCATGTTACGACTT	62.5	49.7	575	0.265	0.777	17	59	52.78
11	CACACCGCCCGTCACTCT	CTTACCATGTTACGACTT	62.8	49.7	572	0.264	0.776	18	59	53.74
12	ACACACCGCCCGTCACTC	CTTACCATGTTACGACTT	63.1	49.7	565	0.261	0.779	16	59	54.60
13	CTTACCATGTTACGACTT	GCACACACCGCCCGTCAC	49.7	65.3	536	0.247	0.804	18	59	55.85
14	CACCGCCCGTCACTCTCC	CTTACCATGTTACGACTT	63.5	49.7	540	0.249	0.789	16	59	51.75
15	CACCGCCCGTCACTCTCC	CACTTACCATGTTACGAC	63.5	51.1	500	0.231	0.846	47	59	53.94
16	ACACCGCCCGTCACTCTC	TACACTTACCATGTTACG	62.5	49.5	497	0.229	0.841	50	59	56.84
17	ACCATGTTACGACTTGCC	CACACCGCCCGTCACTCT	55.8	62.8	546	0.252	0.766	44	59	50.96
18	CACCGCCCGTCACTCTCC	TTACCATGTTACGACTTG	63.5	50.1	530	0.244	0.785	44	59	50.95
19	CGCACACACCGCCCGTCA	CTTACCATGTTACGACTT	66.6	49.7	508	0.234	0.811	19	59	56.71
20	ACCGCCCGTCACTCTCCC	TTACCATGTTACGACTTG	64.5	50.1	483	0.223	0.845	43	59	50.11
21	CACACCGCCCGTCACTCT	TGTTACGACTTGCCTCCC	62.8	57.4	497	0.229	0.817	40	55	47.12
22	ACACCGCCCGTCACTCTC	TGTTACGACTTGCCTCCC	62.5	57.4	497	0.229	0.817	39	54	46.12
23	ACACACCGCCCGTCACTC	TGTTACGACTTGCCTCCC	63.1	57.4	497	0.229	0.817	41	56	48.12
24	ACCGCCCGTCACTCTCCC	TGTTACGACTTGCCTCCC	64.5	57.4	488	0.225	0.830	37	52	44.12
25	CACACCGCCCGTCACTCT	GTTACGACTTGCCTCCCC	62.8	58.4	495	0.228	0.816	39	54	46.12
26	ACACCGCCCGTCACTCTC	GTTACGACTTGCCTCCCC	62.5	58.4	495	0.228	0.816	38	53	45.12
27	ATGTTACGACTTGCCTCC	CACCGCCCGTCACTCTCC	55.2	63.5	517	0.238	0.779	39	55	45.95
28	CCGCCCGTCACTCTCCCC	GTTACGACTTGCCTCCCC	65.5	58.4	479	0.221	0.827	35	50	42.13
29	ACTTACCATGTTACGACT	CACCGCCCGTCACTCTCC	50.8	63.5	505	0.233	0.786	46	59	52.97
30	ACCGCCCGTCACTCTCCC	CATGTTACGACTTGCCTC	64.5	54.2	471	0.217	0.841	39	54	46.09
31	CACCGCCCGTCACTCTCC	GTTACGACTTGCCTCCCC	63.5	58.4	474	0.219	0.829	37	52	44.13
32	ACACCGCCCGTCACTCTC	TTACGACTTGCCTCCCCT	62.5	58.1	483	0.223	0.812	37	51	44.06
33	ACCGCCCGTCACTCTCCC	TTACGACTTGCCTCCCCT	64.5	58.1	473	0.218	0.825	35	49	42.07
34	ACACTTACCATGTTACGA	CACCGCCCGTCACTCTCC	51.1	63.5	496	0.229	0.784	48	59	54.91
35	ACACTTACCATGTTACGA	CACACCGCCCGTCACTCT	51.1	62.8	498	0.230	0.779	50	59	56.83
36	ACACCGCCCGTCACTCTC	GTACACTTACCATGTTAC	62.5	47.9	463	0.214	0.838	51	59	57.66
37	ACACACCGCCCGTCACTC	ACACTTACCATGTTACGA	63.1	51.1	463	0.214	0.838	19	59	57.58
38	ACCGCCCGTCACTCTCCC	CACTTACCATGTTACGAC	64.5	51.1	457	0.211	0.840	46	59	53.07
39	ACCGCCCGTCACTCTCCC	CTTACCATGTTACGACTT	64.5	49.7	491	0.226	0.780	15	59	50.88
40	CCGCCCGTCACTCTCCCC	TTACCATGTTACGACTTG	65.5	50.1	475	0.219	0.775	42	58	49.12
41	CACTTACCATGTTACGAC	CCGCCCGTCACTCTCCCC	51.1	65.5	453	0.209	0.781	45	59	52.09
42	CCGCCCGTCACTCTCCCC	TTACGACTTGCCTCCCCT	65.5	58.1	467	0.215	0.754	34	48	41.07
43	ACACTTACCATGTTACGA	ACCGCCCGTCACTCTCCC	51.1	64.5	453	0.209	0.779	47	59	54.02
44	ACACTTACCATGTTACGA	CCGCCCGTCACTCTCCCC	51.1	65.5	451	0.208	0.780	46	59	53.07
45	CACCGCCCGTCACTCTCC	TTACGACTTGCCTCCCCT	63.5	58.1	462	0.213	0.753	36	50	43.07
46	CACACCGCCCGTCACTCT	TACGACTTGCCTCCCCTT	62.8	58.1	436	0.201	0.729	40	51	44.05
47	ATACCGCGGCCGTTAAAC	GCCTGTTTACCAAAAACA	59.0	51.8	562	0.259	0.541	45	57	51.97
48	ATACCGCGGCCGTTAAAC	CGCCTGTTTACCAAAAAC	59.0	53.3	561	0.259	0.540	46	58	52.97
49	ATACCGCGGCCGTTAAAC	TTACCAAAAACATCGCCT	59.0	52.7	452	0.208	0.633	39	51	46.25
50	ATACCGCGGCCGTTAAAC	CCTGTTTACCAAAAACAT	59.0	48.7	509	0.235	0.534	44	56	51.01
51	ATACCGCGGCCGTTAAAC	CTGTTTACCAAAAACATC	59.0	47.4	499	0.230	0.535	43	55	50.04
52	ATACCGCGGCCGTTAAAC	TGTTTACCAAAAACATCG	59.0	49.3	454	0.209	0.537	42	54	49.26
53	ATACCGCGGCCGTTAAAC	TTTACCAAAAACATCGCC	59.0	52.0	452	0.208	0.535	40	52	47.26
54	ATACCGCGGCCGTTAAAC	TACCAAAAACATCGCCTC	59.0	53.4	442	0.204	0.529	38	50	45.26

Table 3.2: Fifty five primer pairs proposed by *ecoPrimers* to amplify potential barcode markers specific to *Metazoas*. The primers were restricted to strictly match on at least 10% of example sequences (-q 0.1) and to amplify a length of $10-60$ bp. These primer pairs have low taxonomic coverage but highly specific to those taxa, thus can be good candidate for making sets.

They corresponded to 2002 species belonging to 1549 genera. On this data set, we restricted our example set to *metazoa* (NCBI Taxid: 33208). Among these 2044 sequences,

[2] (http://eutils.ncbi.nlm.nih.gov)

1871 sequences belonged to our example set *i.e. metazoa* which corresponded to 1833 species. On this example data set, we used a strict quorum value ($-q = 0.1$) so that strict repeats were present in at-least 10% of example sequences. We further restricted our primer pairs to amplify a length between $10 - 60$ bp. *ecoPrimers* gave us 110 primer pairs.

All of the primer pairs for this example had high value of specificity, moreover most of the primer pairs belonged to same region, thus each primer pair identified the same set of taxa. Due to this reason, the reduced set with graph approach had only few primer pairs. Thus we used the actual 110 primer pairs as input for metaheuristic algorithms. In the Table 3.2, we have shown 55 primer pairs which had been used in the construction of best neighbor set.

Based on these primers the optimal solution sets with both Simulated Annealing and Tabu Search are shown in Tables 3.3 and Table 3.4. From the results we can see that maximum value of coverage achieved by Tabu Search is 0.401 whereas by Simulated Annealing it is 0.405, so both methods gave almost same results. Although the total increase in

Sr.No.	Set B_c	Set B_s	Set Amplified Count	Set well Identified Count	Set Score	Primers in Set
1	0.344	0.847	746	632	0.344	$\{24, 13, 16, 0\}$
2	0.346	0.842	751	632	0.346	$\{24, 13, 42, 11\}$
3	0.381	0.830	827	686	0.381	$\{15, 13, 50, 29\}$
4	0.385	0.834	835	696	0.385	$\{8, 13, 50, 29\}$
5	0.386	0.779	837	652	0.386	$\{54, 13, 50, 42\}$
6	0.395	0.825	856	706	0.395	$\{0, 13, 50, 42\}$
7	0.395	0.823	857	705	0.395	$\{47, 6, 19, 45\}$
8	0.396	0.819	858	703	0.396	$\{47, 17, 19, 31\}$
9	0.398	0.827	862	713	0.398	$\{47, 18, 19, 31, 0\}$
10	0.399	0.828	866	717	0.399	$\{39, 13, 48, 46, 0\}$
11	0.401	0.833	870	725	0.401	$\{40, 48, 1, 12, 19, 31\}$
12	0.405	0.826	878	725	0.405	$\{47, 19, 13, 25, 10, 38\}$
13	0.405	0.835	879	734	0.405	$\{10, 39, 50, 47, 13, 19, 32, 28, 4\}$

Table 3.3: Some sets propositions for primer pairs in table 3.2 using Simulated Annealing heuristics approach.

Sr.No.	Set B_c	Set B_s	Set Amplified Count	Set well Identified Count	Set Score	Primers in Set
1	0.344	0.801	745	597	0.344	$\{43, 41, 47, 1\}$
2	0.344	0.802	746	598	0.344	$\{3, 48, 44, 36\}$
3	0.348	0.795	755	600	0.348	$\{5, 48, 34, 42\}$
4	0.353	0.791	766	606	0.353	$\{27, 10, 47, 23\}$
5	0.354	0.792	768	608	0.354	$\{44, 10, 47, 26\}$
6	0.366	0.856	794	680	0.366	$\{0, 19, 49, 37\}$
7	0.370	0.807	802	647	0.370	$\{35, 19, 49, 42\}$
8	0.379	0.833	822	685	0.379	$\{54, 7, 13, 51\}$
9	0.391	0.834	848	707	0.391	$\{5, 48, 13, 14\}$
10	0.393	0.823	853	702	0.393	$\{22, 48, 13, 34\}$
11	0.395	0.829	856	710	0.395	$\{47, 16, 35, 22, 19\}$
12	0.396	0.824	858	707	0.396	$\{47, 8, 32, 22, 19\}$
13	0.398	0.828	862	714	0.398	$\{13, 47, 30, 33, 1\}$
14	0.399	0.823	865	712	0.399	$\{53, 48, 13, 10, 21\}$
15	0.401	0.832	869	723	0.401	$\{13, 11, 47, 28, 2, 52, 20\}$

Table 3.4: Some sets propositions for primer pairs in table 3.2 using Tabu Search heuristics approach.

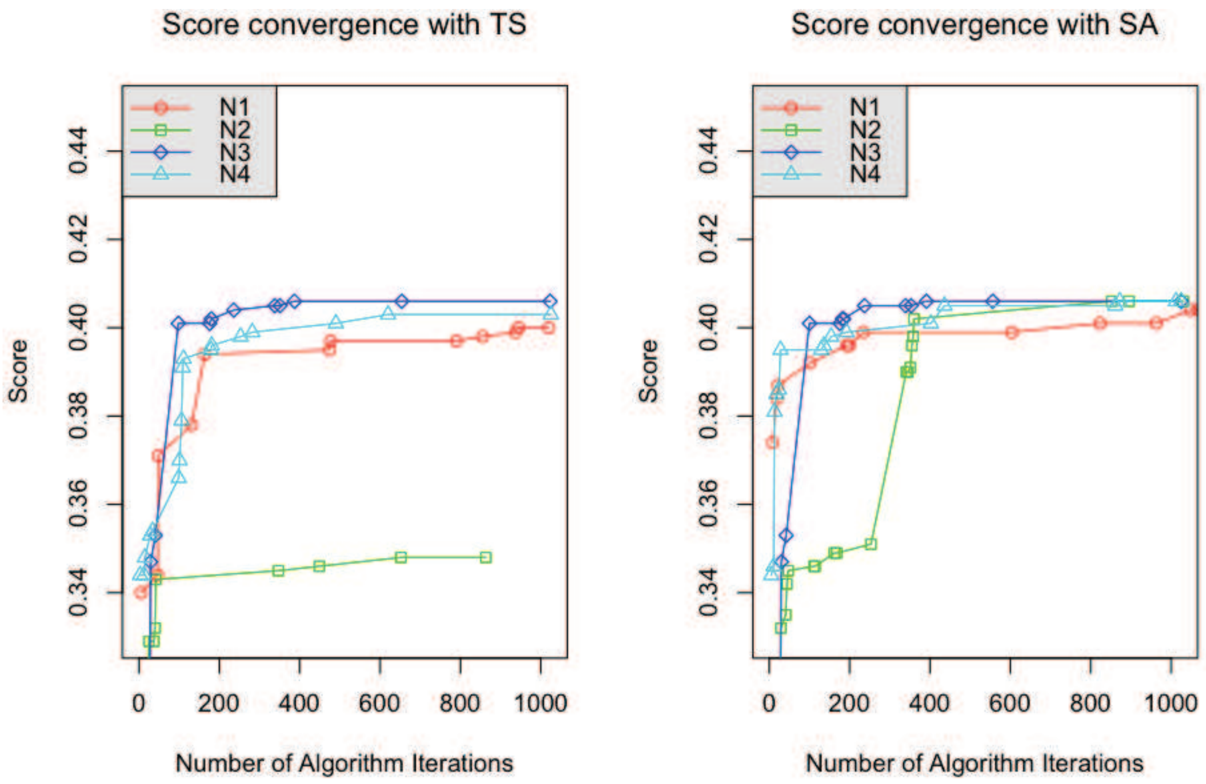


Figure 3.1: This figure shows the convergence of both Tabu Search (left) and Simulated Annealing (right) meta heuristics for four different neighboring criteria: $N1$: *All random*, $N2$: *One with next*, $N3$: *Least contributing with the next* and $N4$: *All random on least contributing*. See section 3.2.5 for a detailed description of these criteria. Other than $N2$ all criteria have same convergence behavior.

coverage by this sets approach is less than 20%, this is because most of the found pairs for *matazoa* belonged to same region so they amplify same set of taxa. The results were further verified by calculating an upper bound of B_c for whole set ($B_c = 0.406$ taking the union of all 110 pairs) which is quite close to the two results. Since we know that both Simulated Annealing and Tabu Search do not block in a local minima so the optimality of our algorithm to find the optimal set is equal to the optimality of these metaheuristics. So we can safely say that the probability of finding the optimal set (if it exists) is fairly high. Figure 3.1 shows the score convergence with respect to algorithm iterations for four different neighboring criteria for both Simulated Annealing and Tabu Search heuristics.

3.3 Can We Avoid PCR ?

Although the technique of PCR has greatly evolved in the last decades, there are inherent problems in PCR like mis-incorporated bases which cannot be avoided. We have already discussed this problem in the first chapter of thesis and based on our data analysis, we will show in the next chapter (chapter 4) that most of the mis-incorporations in a solexa sequencing run were actually introduced during the PCR step. This problem is higher when studies involve the retrieval of DNA from ancient samples because very small quantity of DNA is available and size of preserved fragments are small. We limit this PCR effect of degraded DNA by selecting short markers. In the design of *ecoPrimers*, we ensure this requirement during the pairing of primers. Two primers are allowed to make a pair only if on a particular sequence, they lie within the required amplification length specified by user. However such a condition results in throwing away many primers that lie outside the required amplification length and as a consequence we may lose certain regions with high coverage or high discrimination capacity even without testing them. But, PCR is not the only method of target-enrichment strategies. Other strategies like hybrid capture (Mamanova *et al.*, 2010) of the studied barcode sequence could be an interesting candidate to replace the PCR step. Hybrid capture techniques including Primer Extension Capture (PEC) and Array Based Sequence Capture have been successfully used in many studies especially those involving ancient DNA (Briggs *et al.*, 2009, Burbano *et al.*, 2010), and could be an accurate way to produce sound datasets. Thus in order to avoid the continuous replication of PCR errors and to make full use of conserved regions found by *ecoPrimers* which means, at-least evaluating all the conserved regions to check their identification properties, we propose to use the technique of Primer Extension Capture for studies involving heavily degraded and contaminated DNA instead of direct PCR amplification. The procedure followed in PEC method is explained below.

3.3.1 The Technique Of Primer Extension Capture

The technique of primer extension capture is based on using 5′- biotinylated oligonucleotide primers and a DNA polymerase to capture specific target sequences from an adaptor-ligated DNA library. With this technique, it is possible to isolates specific DNA sequences from complex libraries of highly degraded DNA. The actual procedure is as follows:

First of all a sequence library is prepared by attaching A and B adapter molecules to project specific barcode sequences. Then 5′- biotinylated oligonucleotide primers are added to this sequence library and are allowed to anneal to their target sequences. Then an extension step is performed using *Taq* DNA polymerase which results in the double stranded association between primers and target sequences. At this point some primers may remain unused and in order to remove them spin column purification is performed. Biotinylated primer-target duplexes are captured by streptavidin-coated magnetic beads. The beads are then washed stringently above the melting temperature of the PEC primers, to ensure that templates upon which extension occurred remain associated with the primers. Finally captured and washed targets are eluted from the beads and can be amplified with adaptor priming sites using only 1 or 2 cycles of PCR. The technique is shown in Figure 3.2.

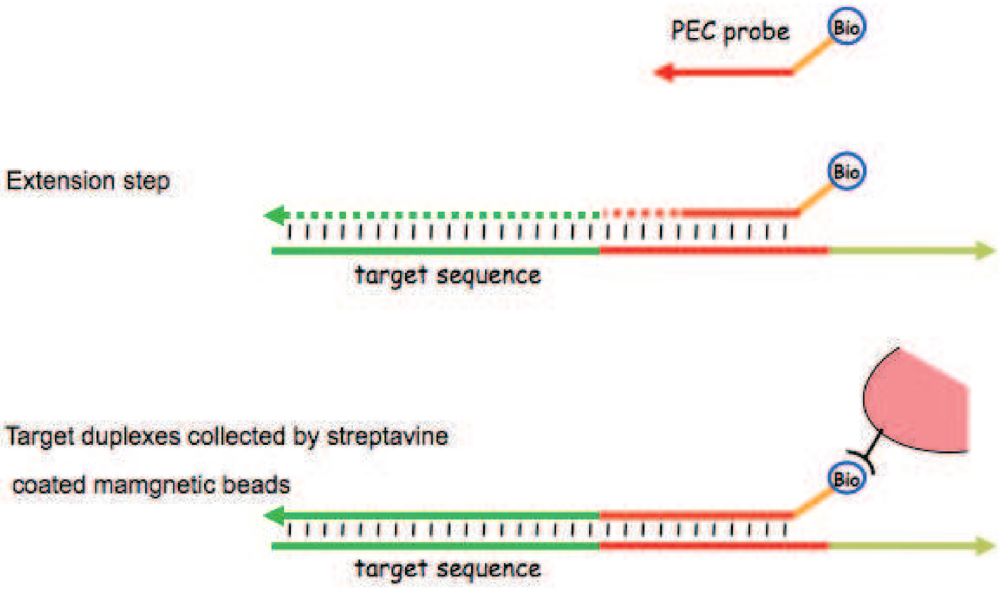


Figure 3.2: 5′ - Biotinylated oligonucleotide primers (PEC primers), extension and collection of target sequences.

The biggest advantage of PEC is that, it greatly reduces sample destruction and sequencing demands relative to direct PCR, thus appropriate for ancient DNA samples. PEC method is simple, quick, sensitive and specific, however this method is not an ideal choice for the capture of very large (*e.g.* a mega bases or more) regions because the sensitivity of capture becomes lower as the number of PEC primers in a multiplex capture reaction increases. This method was originally developed to analyze areas of interest in the Neanderthal nuclear genome (Briggs *et al.*, 2009) however it might also be useful for other types of targeted sequencing of short regions like outside ancient DNA, such as capture of small RNA fragments from an RNA library or capture of 16S (or other loci) diversity from a metagenomic sample *etc*.

3.3.2 PEC Probes Design

In order to use PEC method, we need application specific oligonucleotide primers, in the same way we need primer pairs for PCR. While for PCR experiment, two primers (sense and antisense) are required for building two double stranded DNA molecules from two single stranded DNA molecules by starting extension from opposite ends, in the technique of PEC, only one such primer is required to start the hybridization of a single stranded target DNA molecule. In the context of this study we call these primers as probes in order to avoid the confusion with PCR primers. The PEC probes have the same properties as PCR primers, for example, high taxonomic coverage, high discrimination capacity of amplified region and shorter amplifications length. In order to design such probes, we have developed a program called *ecoProbes* as a small extension of *ecoPrimers*. Thus *ecoPrimers* can design the barcode markers and their associated PCR primers and *ecoProbes* can design PEC probes.

ecoProbes

Our *ecoPrimers* algorithm had the following steps; finding strict repeats, using found strict repeats as patterns and finding the positions of their approximate matches using *Agrep* algorithm, tagging the repeats as *good* or *bad* primers, pairing the primers for each sequence, evaluating quality indices and measuring melting temperature of primers. *ecoProbes* actually uses the implementation of *ecoPrimers* omitting the pairing step. So all of the repeats found can be potential probes depending upon their quality. Since no pairing is required, we have many more probes than PCR primer pairs as no primers are thrown away due to amplification length constraint. Each probe can amplify in both sense and antisense directions, so one probe has two amplifias, one on its right side and one on its left side. Unlike *ecoPrimers*, where quality of a pair is based on B_c and B_s indices (the value B_s strongly depends on the amplification length required), the quality of a probe is based on the amplification length required to attain a certain value of B_s specified by user. This means to say that what length of DNA needs to be sequenced in order to well identify a given k% of taxa. Since a probe can amplify in both direction, the actual sequence of probe can be determined depending on the shortest of sense or antisense amplification lengths. A sample from *ecoProbes* output is shown in Table 3.5 where probes are designed for our *mito* data set, restricting the example set E_s to Vertebrata (NCBI Taxid: 7742) with 80% of $specificity$ value required. For this run, we also tried to show the position of probes and their corresponding amplifia on the sequence with GeneBank *Accession* No: *NC*_013725.

3.4 Conclusion

In this chapter we have presented *ecoPrimers* program for designing optimal barcode markers mostly suitable for *metabarcoding* and generally for both barcoding types. The program has proven to be very efficient for designing primer pairs removing *a priori* on

Sr.No.	*L/RRegion*	*Probe*	*Len*	T_m	B_c	Amplifia
1	L	TTAGATACCCCACTATGC	32	50.70	0.994	519..550; *ctagccgtaaacattgatagaattatacacct*
2	L	GGGTATCTAATCCCAGTT	36	50.5	0.995	*complement*(457..492); *tgtgtcctagctttcgtggggtcgggggtaataaag*
3	L	TGGGATTAGATACCCCAC	37	52.9	0.982	514..550; *tatgcctagccgtaaacattgatagaattatacacct*
4	R	AAACTGGGATTAGATACC	40	48.5	0.996	510..549; *ccactatgcctagccgtaaacattgatagaattatacacc*
5	L	TAGTGGGGTATCTAATCC	41	49.5	0.994	*complement*(457..497); *cagtttgtgtcctagctttcgtggggtcgggggtaataaag*

Table 3.5: Some probes propositions for $B_s \geq 80\%$ for taxid NC_013725 with positions on actual DNA sequence. Column No 2 shows that which of the left or right amplifia is smaller to achieve the 80% value of B_s. Column No. 4 gives the amplification length required to achieve 80% B_s and last column shows amplified sequences with position on actual sequence. "complement" means that complement of this probe should be used.

gene choice by scanning full genome analysis and being able to run on large databases of long sequences. No other available programs proved to be enough efficient. *ecoPrimers* is extended from its basic task of barcode and primer pairs designing to propose optimal sets of short barcode markers that can be used in conjunction to increase the number of identified taxa. This functionality could be very helpful in the context of metabarcoding applications, where long barcode markers cannot be used due to unavoidable constraints of damaged DNA. However many short barcode markers combined can identify as many taxa as a single long barcode marker. We have implemented another extension to *ecoPrimers* for designing probes to be used with PEC technique. Although this functionality cannot be of much help currently because the technique of PEC is in its initial stages, this could be proven quite interesting in the near future with the development of application for example *in Silico* DNA capture like *in Silico* PCR.

3.5 Résumé

Ce chapitre présente le logiciel *ecoPrimers* que l'on a créé pour inférer des barcodes optimaux pour les applications de DNA metabarcoding. Ce logiciel est capable d'utiliser des jeux de données d'apprentissage de grande taille comme l'ensemble des génomes bactériens complètement séquencés. Comme pour le chapitre précédant les résultats principaux sont présentés au travers de notre article sur ce travail. Puis, ils sont étendus de deux façons : une réflexion sur la sélection d'un ensemble minimum d'amorces maximisant le nombre d'espèces identifiées, le développement d'ecoProbe permettant la sélection d'amorces simples pouvant être utilisé avec des techniques de capture d'ADN.

CHAPTER 4

Errors in DNA Sequences

4.1 Introduction

It has been discussed in the first chapter that errors are frequently found in DNA sequences and these errors pose a problem for the correct assessment of biodiversity. In this chapter, I present some preliminary works about these errors, their behavior and some propositions about how to deal with them. Errors cannot just be ignored, they bias MOTUs assignation process as well as species richness and diversity estimations. To reduce this impact we need to identify erroneous reads to de-noise the data in order to provide accurate assessment of biodiversity. By considering different hypotheses, our main work is concerned with checking that at which experimental step most of the errors are introduced into the data.

In order to analyze sequences for learning errors behavior, I worked on a set of simple sequences obtained from the diet analysis of snow leopard (*Uncia uncia*). Snow leopard diet was analyzed using his feces samples. Feces were collected by the field workers of The Snow Leopard Trust[1] in Mongolia during summer 2009. DNA extraction from these samples and PCR protocol used are presented in *ecoPrimers* article in chapter 3. The sequencing was carried out on an Illumina/Solexa Genome Analyzer IIx. The sequence reads were analyzed using *OBITools*.[2] Identical sequences were clustered. Each cluster is called a unique sequence and is weighted by the number of associated sequence reads. Sequences shorter than 10 bp, or containing nucleotides other than A, C, G and T were excluded using the *obigrep* program from *OBITools*. The length of target sequences is between 100 and 108 bp and they belong to mitochondrial *V5* region of the *12S RNA* gene.

Snow leopard's diet analysis is an interesting example for learning errors behavior because

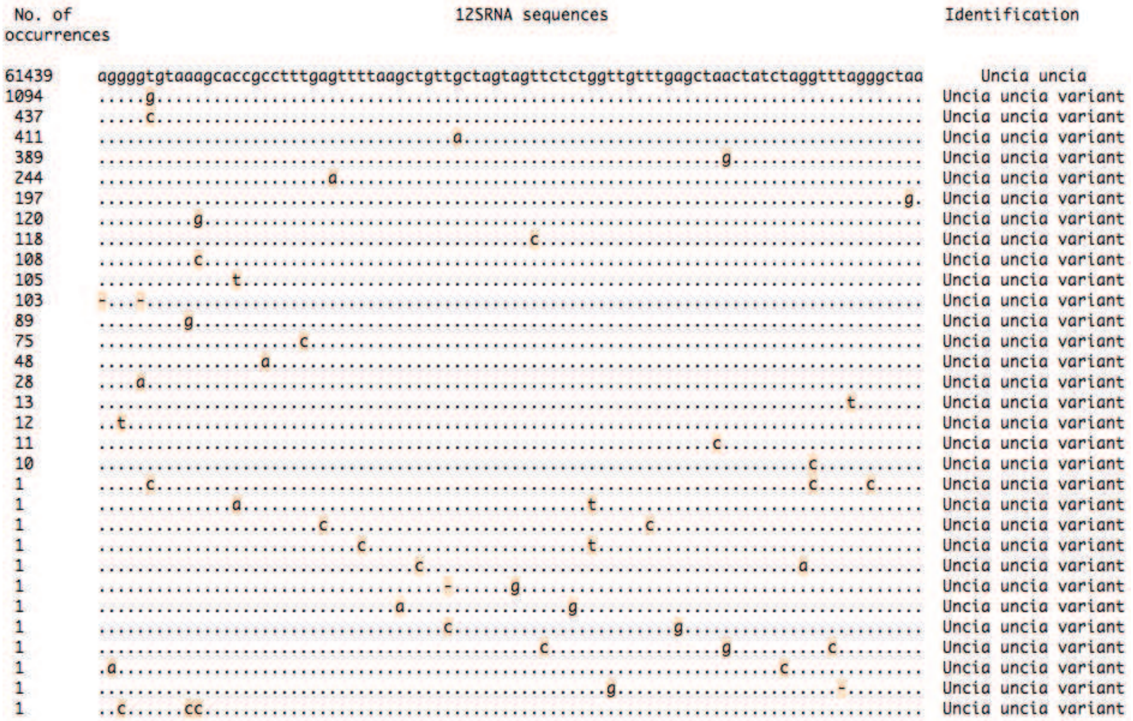
[1] `http://www.snowleopard.org`

[2] `http://www.prabi.grenoble.fr/trac/OBITools`

the diet of this species is simple and consists of mostly mountain goat (*Capra sibrica*) as shown in Table 4 of *ecoPrimers* article in chapter 3. Moreover the reference database is available. We took 10 samples of snow leopard diet which were sequenced from 10 independent PCR runs. In all of the 10 samples, there were two reference sequences which were the true sequences of *Uncia uncia* (UU) and *Capra sibrica* (CS). More stats about this dataset are given in next sections.

4.2 Some Observations About Errors

Using snow leopard samples, we present some observations about the presence of errors in sequences. A simple behavior of errors is shown in figure 4.1 where a true snow leopard sequence is aligned with some of its variants. We can clearly see in this figure that mutations are very frequent and most of the low frequency sequences have higher number of errors.



```
 No. of                                  12SRNA sequences                                        Identification
occurrences

61439   aggggtgtaaagcaccgcctttgagttttaagctgttgctagtagttctctggttgtttgagctaactatctaggtttagggctaa     Uncia uncia
1094    .....g................................................................................   Uncia uncia variant
 437    .....c................................................................................   Uncia uncia variant
 411    .....................................a................................................   Uncia uncia variant
 389    .................................................................g....................   Uncia uncia variant
 244    ........................a.............................................................   Uncia uncia variant
 197    ....................................................................................g.   Uncia uncia variant
 120    ..........g...........................................................................   Uncia uncia variant
 118    .............................................c........................................   Uncia uncia variant
 108    ..........c...........................................................................   Uncia uncia variant
 105    ..............t.......................................................................   Uncia uncia variant
 103    -...-.................................................................................   Uncia uncia variant
 89     .........g............................................................................   Uncia uncia variant
 75     .....................c................................................................   Uncia uncia variant
 48     .................a....................................................................   Uncia uncia variant
 28     ....a.................................................................................   Uncia uncia variant
 13     ..............................................................................t.......   Uncia uncia variant
 12     ..t...................................................................................   Uncia uncia variant
 11     ................................................................c.....................   Uncia uncia variant
 10     ..........................................................................c...........   Uncia uncia variant
 1      .....c....................................................................c.....c.....   Uncia uncia variant
 1      ..............a....................................t..................................   Uncia uncia variant
 1      .......................c.................................c............................   Uncia uncia variant
 1      ...........................c.......................t..................................   Uncia uncia variant
 1      .................................c.......................................a............   Uncia uncia variant
 1      ....................................-......g..........................................   Uncia uncia variant
 1      ...............................a.................g....................................   Uncia uncia variant
 1      ....................................c.......................g.........................   Uncia uncia variant
 1      ..............................................c..................g..........c.........   Uncia uncia variant
 1      .a.....................................................................c..............   Uncia uncia variant
 1      .....................................................g.......................-........   Uncia uncia variant
 1      ..c......cc...........................................................................   Uncia uncia variant
```

Figure 4.1: True *Uncia uncia* sequence is the first sequence with the highest count. This sequence is aligned with some of its variants in order to show that errors in sequences are frequent. Moreover we can see that number of errors is larger for sequences occurring lesser number of times

Such errors are certainly the main reason for artificially elevated microbial diversity estimation (Huse *et al.*, 2010). However one big problem in the field of microbial diversity is the absence of reference database which has led scientists to create the phenomena of "rare biosphere". Taking the advantage of our example data set for which a reference database is also available and there is a simple set of one predator having few prey choices, we can easily reject the hypothesis that low frequency reads stand for real and rare MOTUs. The most possible explanation of low frequency reads is that they are actually erroneous versions of either predator or prey sequences. We can show this with the help of a distance matrix that depicts the distance of each sequence from the reference sequences (in our case true *Uncia uncia* and *Capra sibrica* sequences are the reference sequences). A distance matrix is simply the inverse of a similarity matrix. In a similarity matrix a score of resemblance is calculated for two aligned sequences which shows their similarity, whereas in a distance matrix, a distance shows how distant two aligned sequences are. We estimate the edition distance $D_{i,j}$ between two sequences S_i and S_j by the longest common substring (LCS) (Gusfield, 1997) approach. With this approach, one unit of distance corresponds to one difference between two sequences.

$$D_{i,j} = \max(length(S_i), length(S_j)) - LCS_{i,j} \tag{4.2.1}$$

In figure 4.2 we show a plot based on the distance of all sequences present in one sample from the true *Uncia uncia* sequence.

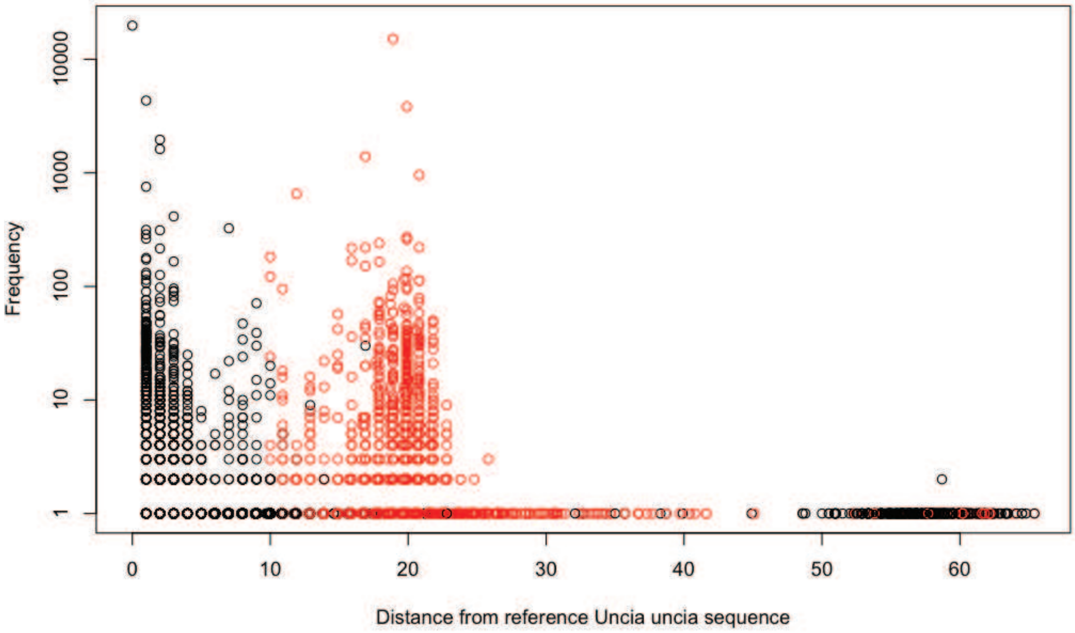


Figure 4.2: Distance of all sequences from true *Uncia uncia* sequence. Each dot corresponds to one sequence. On x-axis is the distance of the sequences from true *Uncia uncia* sequence, whereas on y-axis is the count of occurrence of that sequence. Color is black if the distance of the sequence from true *Uncia uncia* sequence is less than its distance from true *Capra sibrica* sequence otherwise it is red.

We can divide this graph in three groups of sequences. In the first group, there are sequences which are closer to *Uncia uncia* (concentrated black circles on left), second group comprises of sequences closer to *Capra sibrica* (concentrated red circles on right) and the third group consists of sequences which are in the middle of two groups. These sequences are at equal distance from both reference sequences. We can see a lot of singletons here as well. We know that there are only two true sequences, one of *Uncia uncia* and the other of *Capra sibrica* (both highest count sequences with black and red circles), but we observe a lot of other high count sequences which are at a few nucleotide distance from the reference sequences. In order to explain what these sequences stand for, we need to find the actual distinct groups in our data. The simplest explanation of the sequences lying in this group of figure 4.2 could be that these are chimeric products and the singletons are possibly sequencing errors.

In order to find the actual distinct groups in our sample, we projected the whole of our distance matrix into $n-1$ dimensional space using principal coordinate analysis (PCO) implemented in the *ade*4 *R* package (Dray and Dufour, 2007). It is possible to distinguish *Uncia uncia* and *Capra sibrica* sequences using only the 15 first 5′ bases of the marker or the 15 last 3′ bases. We used this property to simply identify chimera sequences. Thus we classified a sequence as UU or CS if both its ends were strictly identical to *Uncia uncia* or *Capra siberica* sequence respectively. Chimeric sequences were tagged CSUU or UUCS depending if they start or end by one or the other reference sequence. All sequences not matching perfectly from the beginning or end to these references are tagged XX. The projection following to the two first axis of the PCO is shown in figure 4.3 and dots are colored according to previously described classification. With such a classification we can see that we have two big groups of UU (light blue) and CS (light green) sequences. The sequences represented by UUCS (pink) and CSUU (dark blue) are the chimeric products.

From these two plots, based on a simple example of snow leopard diet, we see that a huge amount of erroneous sequences are produced, most of which are close variants of predator or prey sequences. However it is important to understand that at which steps of experiments these noisy reads are produced and which DNA regions are more susceptible to errors, so that techniques can be developed to deal with errors and clean data can be made available.

4.3 Questions And Hypotheses

As we have seen in preceding section that during an experiment many erroneous sequences are generated, it could be quite useful to know during which experimental steps

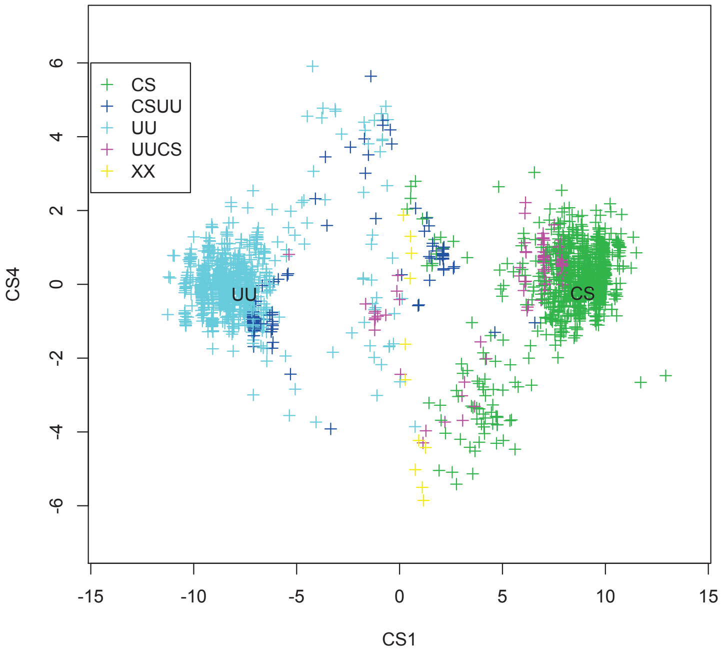


Figure 4.3: Similarity projection using *Principal Coordinate Analysis* technique to show the similar groups of sequences in snow leopard diet analysis. As clearly visible, sequences in two big groups are similar to *Uncia uncia* (light blue) and *Capra sibrica* (green). According to our model sequences in groups UUCS and CSUU are chimeric products.

these errors are introduced. To find an answer to this question we need to consider all steps of an experiment where an error may be introduced, they include: 1) initial replication from DNA template, 2) PCR amplification and, 3) sequencing. Moreover if ancient or degraded DNA is used as input for the experiment then errors due to DNA degradation like depurination will be amplified during the initial step of the PCR.

In order to be sure that we analyze erroneous sequences only, from each of 10 samples, we selected all the sequences which had a single nucleotide difference with one or both of the reference sequences. *obipcrerror* program from *OBITools* was used to identify sequences with one difference from reference sequences and to characterize errors. For each retrieved sequence we recorded the position of the error, its type: insertion, deletion or substitution and its sub-type (*e.g.* $A \rightarrow C$ for a substitution or deletion of a A from an homopolymer of length 4). Further we looked for double errors corresponding to the combination of single errors. With this error data set we are interested to answer the following questions:

- Are all DNA sites equally probable to suffer from errors?
- Are all experimental steps equally probable to generate errors?

In each of the 10 samples, some of single base errors are found only once and some occur multiple times. The proportion of single read errors (*i.e.* cluster of cardinality one) with respect to multiple reads errors (*i.e.* cluster of cardinality greater than 1) is given in figure 4.4. It is noteworthy that most of the errors occur higher number of times.

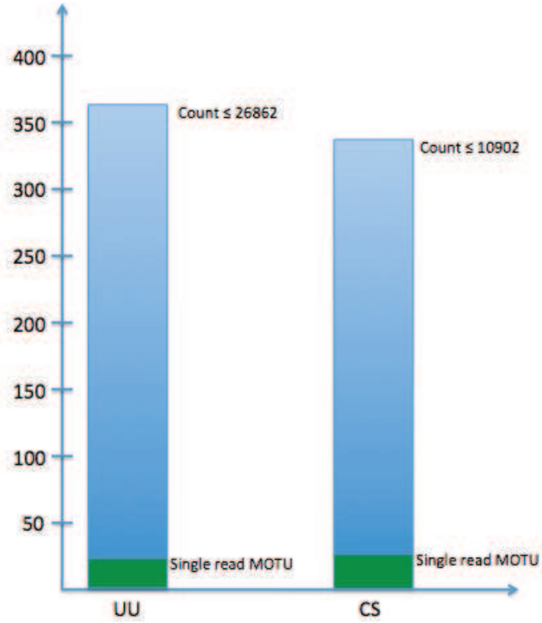


Figure 4.4: Distribution of erroneous reads between clusters of size one and greater than one. The frequency of single read clusters is very low as compared to total reads.

Frequency of sequences in each of 10 PCR samples and the frequency of those sequences from these samples which are at a distance 1 from both *UU* and *CS* true sequences are shown in table 4.1. Frequencies are expressed in number of reads.

Sample	Total Sequences	No of Sequences (d=1) from UU	No of Sequences (d=1) from CS
S01	186718	19367	8580
S02	111992	26260	1129
S03	142607	29050	1848
S04	151251	25300	2769
S05	109782	15786	3408
S06	62468	13222	127
S07	122684	21825	853
S08	87396	11072	8229
S09	180816	26424	4212
S10	150110	24453	2917

Table 4.1: Frequencies of sequences which are at a distance of 1 nucleotide from *Uncia uncia* and *Capra sibrica* reference sequences for all 10 sample. Frequencies are expressed in number of sequence reads.

4.4 Some Important Error Properties

4.4.1 Probability Of Errors Is Not Uniform

We can split the sources of error (*i.e.* initial replication from degraded template, PCR amplification, and sequencing) in to two different classes. Initial replication and sequencing belong to the first class and both of them can be assimilated to a single replication process. For this class all errors are independent and relative frequency of each error can be considered as an estimation of the occurrence probability of this error. The second class corresponds to errors occurring during PCR amplification. When an error occurs during one PCR cycle, it is amplified in the following cycles and sooner an error occurs, more reads we get at the end of the PCR. Thus for this class of errors, more probable is an error, sooner it occurs during the PCR, generating more reads at the end of PCR. So we can postulate that the number of reads of an error over ten PCR is a proxy to its probability of occurrence. Figure 4.5 shows highly heterogeneous number of reads at each position of the marker for both the reference species. From this figure, some DNA sites seem to be more probable to suffer from errors than others indicating a non-uniform and a highly biased error process.

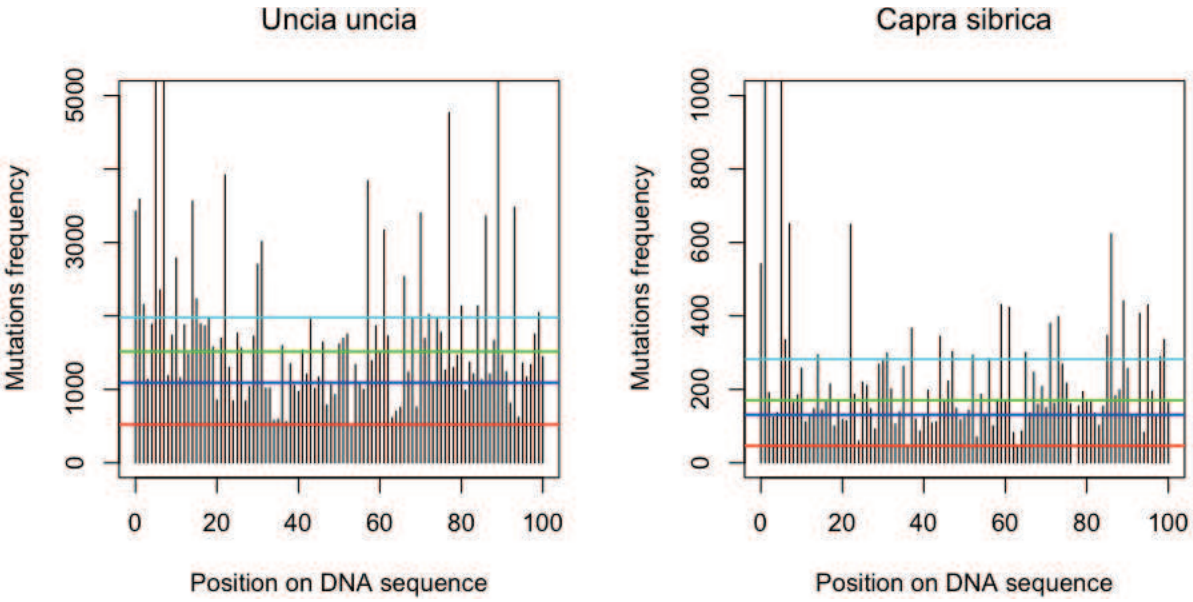


Figure 4.5: Number of reads with one error at each position on DNA sequence for all 10 samples. The horizontal colored lines show the 1st quantile(red), median(blue), 3rd quantile(green) and 4th quantile(cyan).

A second way to assess that error process is highly biased is to compare errors observed in all the ten independent PCR samples. For both *Uncia uncia* and *Capra sibrica*, we have classified all observed errors according to their position, type and subtype. For each independent PCR we ranked these error classes according to their frequency. Ordering obtained for each PCR was compared using Kendall-Tau rank correlation test. For each of the 45 pairs of PCR sample, after bonferoni correction for multiple tests, p-value of the Kendall-Tau test is estimated to 0 for both species, demonstrating a high similarity between error patterns. This consolidates our impression of a highly biased error process. The correlation diagram and p-values for *Uncia uncia* and for *Capra sibrica* are shown in figure 4.6 and 4.7 depicting a positive correlation between any two PCR samples.

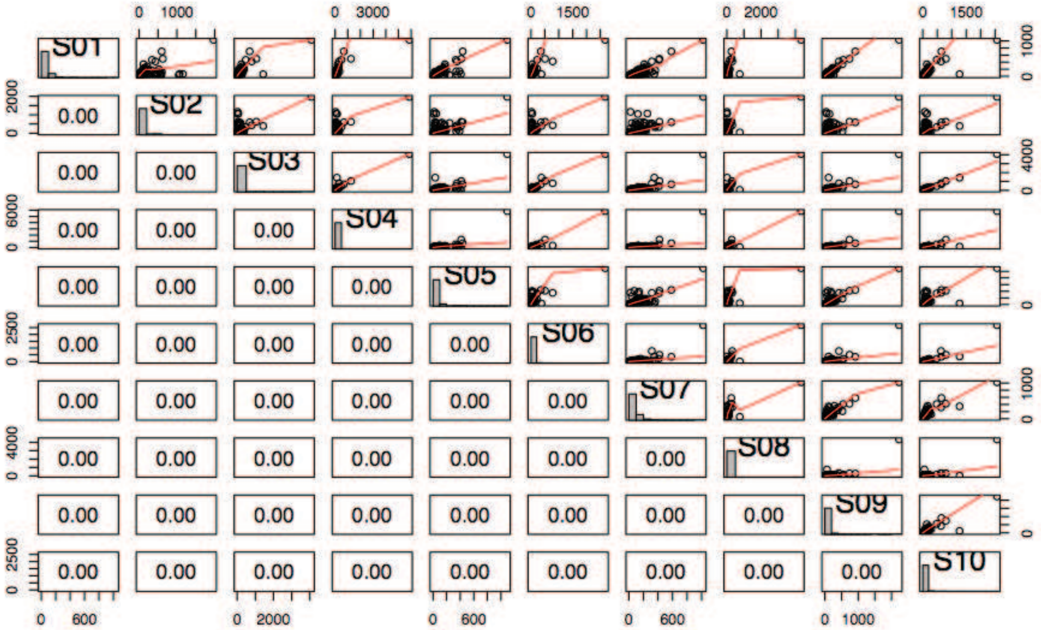


Figure 4.6: Kendall Tau correlation test on 10 samples for UU sequences. Upper triangle shows the correlation graphs, clearly a positive correlation exists between all pairs of samples. Lower triangle shows the $p - values$ of all pairs.

Since a lot of mutations were common between all the samples, so we joined the ten samples in order to have one global error pattern with their associated frequencies for each species. A total of 367 unique sequences corresponding to single base errors were identified for *Uncia uncia* and 340 for *Capra sibrica*. Table 4.2 recapitulates all types and sub-types of mutations globally observed over ten samples. Most of the reads with single base error corresponded to substitutions. Reads corresponding to transitions and transversions occurred at similar frequencies but more than half of the transversion reads were $t \rightarrow g$ substitutions.

4.4.2 Errors Occur Preferentially During PCR Amplification

Previously we divided errors in to two classes and we said that if errors occur preferentially during PCR initiation or sequencing step, it implies that errors occur independently

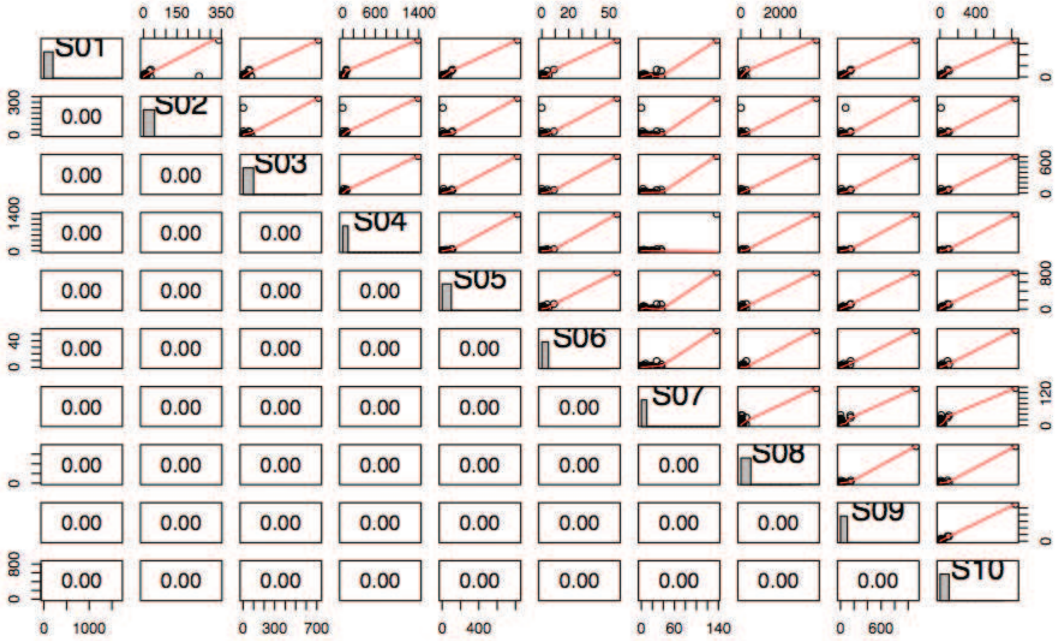


Figure 4.7: Kendall Tau correlation test on 10 samples for *CS* sequences. Upper triangle shows the correlation graphs, clearly a positive correlation exists between all pairs of samples. Lower triangle shows the $p - values$ of all pairs.

Mutations	Sub Types	UU	CS
Insertions		152	41
Deletions		897	205
Transitions	$c \to t$	17740	1159
	$a \to g$	23315	4154
	$g \to a$	29441	2129
	$t \to c$	40817	5981
	Total	111313	13423
Transversion	$a \to t$	2235	387
	$t \to a$	3578	729
	$c \to a$	4649	496
	$c \to g$	5288	658
	$g \to c$	9252	788
	$a \to c$	10892	1789
	$g \to t$	14593	1555
	$t \to g$	49910	14001
	Total	100397	20403

Table 4.2: Different types of single base mutations and their frequencies found in both species.

and relative frequencies of corresponding reads is an estimation of the occurrence probability. Under this hypotheses, we can predict occurrence frequencies of sequences with two errors from the product of each single error frequency (by two errors, we mean a sequence that is at a distance of two base pairs from reference sequences). On the other hand, if errors occur preferentially during PCR amplification, errors are not independent. This is because when a first error occurs, it is amplified and then a second error may occurs during one of the following PCR cycles, eventually affecting a sequence carrying the previous error. Thus for such a situation, read frequencies are not a direct estimation of occurrence probabilities and it is not possible to estimate frequencies of sequences with two errors.

Frequency Estimation Of Sequences With Two Errors

If we suppose that most of the single base errors that we observed for both reference species preferentially occurred during PCR initiation or sequencing step, then frequencies of sequences with two errors can be estimated as following :

Suppose two errors m_1 and m_2 occur at relative frequencies of $\mathcal{F}_1$ and $\mathcal{F}_2$ respectively. If $\mathcal{N}$ is the total number of sequences in all of 10 PCR samples, $\mathcal{N}_{uu0}$ is the number of true *Uncia uncia* sequences and $\mathcal{N}_{cs0}$ is the true *Capra sibrica* sequences. Then we can calculate $\mathcal{N}_{uu}$ the possible number of sequences belonging to UU (all UU sequences including true and erroneous versions in total $\mathcal{N}$ sequences) by using following equation 4.4.1.

$$\mathcal{N}_{uu} = \mathcal{N} \times \frac{\mathcal{N}_{uu0}}{(\mathcal{N}_{uu0} + \mathcal{N}_{cs0})} \tag{4.4.1}$$

And similarly we can calculate the total number of sequences $\mathcal{N}_{cs}$ belonging to *Capra sibrica* by a similar equation 4.4.2

$$\mathcal{N}_{cs} = \mathcal{N} \times \frac{\mathcal{N}_{cs0}}{(\mathcal{N}_{uu0} + \mathcal{N}_{cs0})} \tag{4.4.2}$$

Once we have the approximate number of total sequences belonging to both species, we can calculate the total number of erroneous sequences $\mathcal{N}_{(uu)m}$ belonging to *Uncia uncia* by:

$$\mathcal{N}_{(uu)m} = \mathcal{N}_{uu} - \mathcal{N}_{uu0} \tag{4.4.3}$$

And total number of erroneous sequences $\mathcal{N}_{(cs)m}$ belonging to *Capra sibrica* by:

$$\mathcal{N}_{(cs)m} = \mathcal{N}_{cs} - \mathcal{N}_{cs0} \tag{4.4.4}$$

If $\mathcal{N}_{(uu)1}$ is the total number of single base errors of *Uncia uncia*, then the frequency of a double error resulting from the combination of two single base *Uncia uncia* errors can be calculated using following equation:

$$\mathcal{F}_{uu(12)} = \{\frac{\mathcal{F}_{uu(1)}}{\mathcal{N}_{uu}} \times \frac{\mathcal{F}_{uu(2)}}{\mathcal{N}_{uu}}\} \times (\mathcal{N}_{(uu)m} - \mathcal{N}_{(uu)1}) \tag{4.4.5}$$

And given $\mathcal{N}_{(cs)1}$ is the total number of single base errors of *Capra sibrica*, we can calculate the frequency of double error resulting from the combination of two single base errors with the following equation:

$$\mathcal{F}_{cs(12)} = \{\frac{\mathcal{F}_{cs(1)}}{\mathcal{N}_{cs}} \times \frac{\mathcal{F}_{cs(2)}}{\mathcal{N}_{cs}}\} \times (\mathcal{N}_{(cs)m} - \mathcal{N}_{(cs)1}) \qquad (4.4.6)$$

To test above hypothesis experimentally, we combined two single base observed errors and looked for corresponding double error sequences in our data set. We added up the frequencies of all found instances of this sequence in all of ten samples to get total count of occurrence of double error sequence. Using the counts of single errors and equations 4.4.5 and 4.4.6 we calculated the theoretical value of double error frequencies. Repeating this process for all possible pairs of single errors, we could get experimental and theoretical values of double errors count/frequency. In almost all cases the observed values were much greater than theoretical values implying that double errors could not have occurred during PCR initiation or sequencing steps. To further strengthen this point we ran *Mann-Whiteny U* test on the two vectors of observed and theoretically calculated values of double mutation counts which gave a $p - value = 0$ implying that both theoretical and observed double mutation frequencies are not comparable.

4.4.3 The Error Pattern Is Similar Between *Uncia uncia* And *Capra siberica*

In the previous sections we have discussed that any two PCR samples either of *Uncia uncia* or *Capra siberica* follow the same error patters. We have seen previously that error frequencies from one PCR to the second PCR are positively correlated. However it is also important to check that, does both predator and prey species follow the same error pattern or not?, as both species are amplified in the same PCR. Thus we compared the error patterns of *Uncia uncia* and *Capra siberica* in a similar way previously used for comparing independent PCR. Figure 4.8 shows the correspondence between the two error patterns, where we plot position wise error types frequency. A $Kendall - Tau$ test gave a $p_{value} = 0$. There exist seventeen differences between *Uncia uncia* and *Capra siberica* sequences. As we have previously shown, error pattern is highly biased in position and in class of errors for one version of marker. From the correlation shown in figure 4.8, it is evident that the same errors occur with same frequency at same positions on two different versions of marker. This is not really surprising if we suppose that the bias is related to chemical and physical constraints and that all versions of the marker are highly similar. But since errors are similar after some error accumulation, erroneous sequences originating from both the species share common characteristics. This creates a kind of attraction point in the space of sequences. Such groups are also visible in figure 4.3. Standard classification methods unaware of this behavior would create some extra classes for these groups leading to an over estimation of the number of taxa.

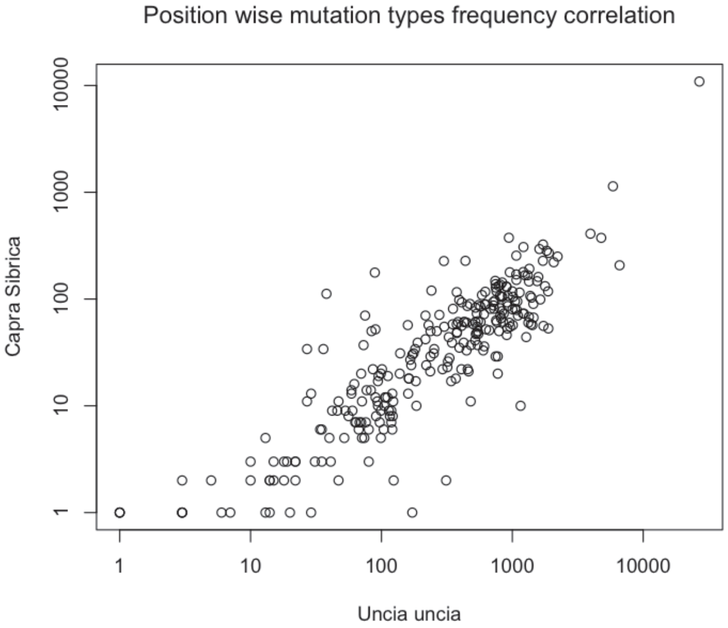


Figure 4.8: Position wise type mutations frequencies graph between UU and CS sequences. It is clear that a strong positive correlation exists between them implying not only the existence of preferential mutation sites but also the most likely step where errors are introduced is PCR.

4.5 Dealing with PCR errors

Since most of the errors seem to be probably generated during the process of PCR, we devised an algorithm to deal with PCR errors and de-noise the sequence data. In this algorithm a directed graph $G(V, R)$ was built, where the set V of vertices is the set of unique sequences M_i from one PCR and R is a directed relation such that two unique sequences M_i and M_j are linked if the distance d_e between two corresponding sequences is equal to one nucleotide. As each unique sequence is weighted by the count of associated reads, edges are directed from the highest weight W to the lowest one in G. Such a graph forms a network between unique sequences where M_i with higher weight is placed above M_j. In such a graph each connected component is a directed acyclic graph (DAG). In this graph each unique sequences is classified either as **'head' (H)**, **'internal' (I)** or **'singleton' (S)** where **'head'** is the root of a DAG, **'singleton'** corresponds to DAG of size one and **'internal'** are the all other nodes. This algorithm was implemented in *obiclean*[3] program.

[3] https://www.grenoble.prabi.fr/OBITools

In this algorithm all unique sequences which are **'head'** or **'singleton'** could be considered as the real sequences.

obiclean algorithm was used by Sophie Prud'homme, during her masters for studying plants diversity in Roche Noire valley (French Alps). The aim of this project was to evaluate relative effect of sampling, DNA extraction and PCR amplification on the variability of metabarcoding results. Thus *obiclean* was used to eliminate the artifactual sequences produced by different steps of the experimental procedure and a metabarcoding approach was used on the de-noised data set to see if the amplification of a barcode in a DNA mixture from temperate region soil could allow a realistic view of the current plant biodiversity. The distribution of the identified MOTU among the different sites was compared to distribution of the corresponding species in the botanical relevés in order to evaluate the accuracy and the validity of the metabarcoding approach. The results obtained from de-noised data set with metabarcoding approach were comparable to botanical relevés statistics which validates our *obiclean* program.

4.6 Conclusion

Environmental metabarcoding presents several important advantages compared to traditional biodiversity assessment methods. According to the used barcode, this method could allow diversity studies on weakly observable organisms, as organisms living in soil or in sediments. However as one of the final aim of environmental metabarcoding is to use it to study diversity of all types of organisms, including the most poorly known taxa, it is necessary to be extremely confident of the obtained list of MOTUs. Experimental noise in PCR and sequencing steps has a strong impact on artificially elevated diversity estimates. In order to deal with this problem, in this chapter we performed a preliminary analysis on a simple example of snow leopard diet where PCR product was sequenced on a Illumina/Solexa Genome Analyzer, and it seems that PCR amplification is highly biased and most of the mutations are generated during this step. We tried to run the same protocol on another data set obtained from 454 sequencer in order to see if the same type of mutation behavior is found with this sequencer. This data set corresponded to the controlled diet analysis of sheep that was kept in a farm and provided with only two plants species Ray Grass (*Lolium perenne*) and Luzerne (*Medicago sativa*). The sheep were provided with both plants in different proportions. The diet was analyzed with two methods, using feces samples and using direct stomach contents and *g*/*h* region was chosen as a potential barcode region in order to determine the proportion of both plants in sheep diet. With 8 sheep used in this experiment and different proportions of plants in diet a total of 96 PCR were performed. We chose 20 PCR samples from this data set and run the same protocol. It was observed that some sites are again more probable to suffer from mutations, however the results of other tests didn't exactly match with the results obtained with snow leopard diet. This is mainly because 454 sequencer provided less number of reads. In each of 20 samples a small number of sequences were present and even smaller number of those which were at a distance of 1 nucleotide from the two reference sequences. Thus enough data was not available to be certain about the validity of mutations behavior in a 454 sequencing run.

Although our preliminary analysis leads us towards the conclusion that most of the errors do not occur during sequencing and seem to be occurring during PCR amplification, still the odd $T \rightarrow G$ transversions are somewhat difficult to explain. Thus it is important to perform the same protocol on more data sets from Solexa sequencer in order to see if these type of transversion occurred by chance or we find the same behavior throughout the other datasets. The overall observations about occurrence of mutations during PCR are quite valuable in the context that the common practice of integrating the product of all PCR samples in order to remove noise is not an elegant solution. It is really important to see if all the PCR samples from the same sampling point show the similar behavior or not. In our example we observed a similar behavior of mutations in all 10 PCR samples. However if one PCR does not correlate with others, it is important to remove that PCR and then denoise the remaining PCR samples in order to have realistic views of diversity.

4.7 Résumé

Ce chapitre présente une série de résultats préliminaires concernant l'analyser de données de séquençage afin d'identifier les sources potentielles d'erreurs. Les résultats présentés montrent que la plupart des erreurs sont générées pendant l'amplification PCR et non pendant le séquençage comme cela était principalement postulé. Je termine ce chapitre en suggérant la réalisation d'analyses similaires sur autres données afin d'étendre la pertinence de nos observations.

CHAPTER 5

Discussion

The precise knowledge of species distribution is a key step in biodiversity studies and in conservation biology. However, species identification can be extremely difficult in many environments, specific life stages and in populations at very low density. This study presents DNA metabarcoding as a suitable method available today for species inventory. It is an essential tool for field identification, and for exposing further layers of biodiversity beyond that which is revealed by traditional methods. By using suitable barcode markers in diversity studies, the species inventory can become more certain, more exploratory and more revealing.

5.1 Evaluation Of Barcode Markers

With the emergence of the concept of metabarcoding, the constraints on the use of ideal barcode loci are relaxed. While the classical barcoding requires to use standard markers, ecologists prefer to use any suitable marker adapted to their study. In this context, the first important challenge of metabarcoding is, the selection of the best DNA region(s) to be used as barcode considering the aims of a study. For this purpose, the *in silico* approach (Ficetola *et al.*, 2010) along with two quality indices B_c and B_s can be used for the identification of the most suitable markers *a priori*. The two formal measures B_c and B_s are formalized using taxonomic information and can rank different barcode markers according to their amplification and taxa discrimination capacity. The comparison of different barcode markers is very important in metabarcoding applications particularly. This is because more than one taxa are present in an environmental sample and thus it is important to use highly specific markers in order to avoid the over-amplification of rare species with low number of mismatches. In such a situation *a priori* knowledge of primers quality can be a great help. This approach has been successfully used by (Bellemain *et al.*, 2010) for the analysis of ITS primers. This study showed that some ITS primers when used with higher number of mismatches potentially introduce bias during PCR amplification and that different primer combinations or different parts of the ITS region should be analyzed in parallel, or alternative ITS primers should be searched for.

5.2 Design Of Barcode Markers

The design of universal barcode markers with high resolution capacity is no doubt an important task in DNA barcoding and it can help in broad scale analysis of life on earth. However it has been argued by some authors that finding a minimum amount of gene sequence data that accurately represents the whole genome of all plants or animals is an impossible task (Rubinoff *et al.*, 2006). This is true, because, even *COI* gene that has been considered a universal barcode marker for animals does not evolve enough in some groups like Cnidaria and has much less *COI* divergence in this phylum as compared to other phylums.

Nevertheless, this region has long been used in animal molecular systematics, initially there was no compelling *a priori* reason to focus on this specific gene among the 13 mitochondrial Protein Coding Genes and 2 ribosomal RNA genes (16S and 12S) for DNA barcoding. Though *COI* fragment does have the advantage of being flanked by two highly conserved "universal " primer sites which has been helpful for automating the collection of DNA barcodes from a diverse range of organisms, but the long length of this region is a big hindrance to its applicability in environmental studies. It is thus necessary to search for alternative DNA barcodes to avoid an exclusive reliance on *COI*. One more important reason to look for new barcode markers is that, in the context of DNA metabarcoding, we have changed the definition of barcode quality, so the standard markers do not fit well in this new definition. The *sensu stricto* barcoding approach promotes the barcode regions which are highly discriminant at species level. In order to achieve this high resolution, length of markers has to be increased. However in metabarcoding, ecologists prefer to amplify as many individuals as possible and then discriminate among most of them to any taxonomic level if resolution is not sufficient for species level discrimination. Thus for metabarcoding applications, B_c is more important than B_s and the shortest possible length is a great concern.

For this purpose, *ecoPrimers* program is an efficient and robust application. It is based on a simple syntactic approach for primer design and has more efficient computation algorithms. The full integration of taxonomy and a large number of parameters are a gateway for designing well-adapted barcode markers for any study (these parameters are discussed in detail in the discussion of *ecoPrimers* article in chapter 3). The biggest advantage of *ecoPrimers* over other programs is that it looks for conserved regions using simple yet efficient algorithm and thus it is able to scan huge databases. However, some readers may argue that primer selection criteria does not fit well with the requirements of reliable PCR amplifications. This is somehow true because our primer selection criteria does not take into account the properties like, checking for self complimentarily, adjusting T_m of both forward and reverse primer *etc*. Although these properties are quite important and can strongly affect PCR amplification product, our main aim was to find out the regions which are universal and sufficiently discriminant and to be able to do so on large sequence databases by scanning the full genome (like fully sequenced bacterial genomes). If we try to ensure desired amplification properties by using accurate estimates of melting temperature, the computation cost will increase. This is the main reason why most of the primer design algorithms focusing on thermodynamics properties for selecting primer pairs, look for such primers either in a single sequence or in well known sets of gene sequences or in small number of pre-aligned sequences. Since *ecoPrimers* uses all the strict repeats which are present in strict quorum $q\%$ of sequences for primer design, so we have the advantage of having a large number of primer paris belonging to same region. In such a situation, one can easily select the pair which has a good balance between B_c, B_s and T_m.

Using this program, we have identified a new short and efficient barcode marker called $12S - V5$. This barcode marker is short enough to be easily sequenced for environmental applications and has high values of B_c and B_s indices. One important point to notice is, that with *ecoPrimers*, most of the barcode markers selected lie on ribosomal RNA genes and only few exceptions of protein coding genes were found when run on chloroplast DNA sequences. One of the reason of not finding protein coding regions for vertebrates with *ecoPrimers* could be the amplification length constraint that was set to be between 50 and 150 nucleotides, in order to have shortest possible regions. Nevertheless, $12S - V5$ barcode marker due to its short amplification length and high amplification and resolution capacity, seems an ideal choice for studies involving amplification from a degraded DNA. Due to these properties this primer pair has already been used in three different environmental studies involving, carnivores diet analysis and studies on soil DNA for obtaining information on past and present ecosystems Epp *et al.*, Shehzad *et al.*(submitted articles, see annexes for the manuscript). Qualities of this marker, allows to envisage its use for routine analysis and that led us to deposit a patent in collaboration with a company protecting its use for commercial purpose. Due to different aims of metabarcoding, the barcode markers mostly suitable for such studies are not very general and thus it is possible that a single barcode marker cannot successfully identify all of the individuals present in an environmental sample. Thus we tried to develop a technique where the most optimal barcode markers can be efficiently chosen from a given pool of markers in order to use a set of primer pairs in a single PCR for increasing the B_c index. However our results based on *Metazoas* (detailed in chapter 3) do not reveal wether the use of sets of barcode markers is really an efficient strategy. The most apparent reason for a low increase in the B_c seemed to be due to all primer pairs belonging to same region and thus not maximizing the union of this index. As a second example, I tried to design primer pairs for Nematodes. Since primer design is this group is complicated, even with very relaxed parameters on *ecoPrimers*, only few primer pairs were found with high number of mismatches. And thus sets approach was not really useable in this case.

In future, if we wants to work further on sets approach, it should be used on exhaustive data sets and see if the upper bound on B_c can really be increased. If the upper bound on B_c is sufficiently large to be accepted then more efficient techniques can be used to further reduce the solution space or simply reduce the set cover problem to a simple polynomial problem so that all the sets can be output in order to make sure that the optimal set is not missed by metaheuristic. However, if the upper bound is always low as it was in the case of *metazoas*, then it could be interesting to use other target enrichment techniques like using probes and DNA capture techniques or even global sequencing, instead of using PCR amplification with specific markers and sequencing of only that specific region.

5.3 Analysis Of Sequence Data

The third important thing to consider was error sources in DNA sequences. The precise boundaries of errors origin cannot be detected due to a large number of parameters that needed to be estimated. DNA degradation, sampling biases, extraction biases and PCR artifacts and finally sequencing errors all play in the accumulation of errors. However in the first step it could be important to prove that most of the odd sequences observed in any dataset are not rare taxa and thus it is important to be careful in diversity estimations. This was achieved using a simple example of snow leopard diet. So if some rare species existed, we should have found much more than 2 species in our amplified product, but actually we observed only two species and some close or distant variants of both reference sequences and only very few (8 sequences, almost 0.3% of total sequences) were found which did not resemble to any of the reference species. Among them 2 sequences were identified as belonging to *Okapia johnstoni* and 6 were identified as *Bos taurus*. There could be two reasons to have these 6 sequences which were different. One is that they may be chimeric sequences but they are so much altered that they resemble other species.

This makes sense because *Okapia johnstoni* is an african species and snow leopard feces sampling was carried out in Mongolia, and *Bos taurus* is a domestic cow. The second reason could be that these sequences are assigned to wrong species in *Genbank*. In any case no rare diversity exists in our dataset, but, data analysis conducted in a similar way for micro-organisms biodiversity estimation would say the opposite. Thus we can safely assume that a small portion of microbial diversity may constitute the rare biosphere but not the whole part and we need to develop techniques in order to assign taxa in the absence of a reference database or a probabilistic model that can give us the probability of one sequence being generated from a given true sequence.

When I started working on the analysis of sequence data, my assumption was that most of the erroneous reads come from sequencing errors. And with this assumption, I developed an HMM model for 454 sequencer to find resemblance between true sequence and its low frequency sequences. For this purpose I made use of the quality scores provided with sequence data. Using Baye's theorem and forward algorithm, I tried to calculate the probability that each low frequency read was generated from a high frequency true sequence. During this analysis we observed two things, first even the low frequency reads had high quality scores. And for some data sets two high frequency sequences were found which had the difference of only one nucleotide making it difficult to decide which is the true sequence. These two issues changed our assumption that errors are mainly generated during sequencing and we started looking for other possibilities that may be it is the PCR step which is most biased. My results in this regard are very preliminary and based on some correlation among different properties of errors. I have shown that most of the errors seem to be generated during PCR amplification. Nevertheless, the same correlations needed to be calculated for a huge amount of other data sets in order to see if same behavior of errors is observed or not, and if same behavior is observed, it is important to design strategies to lower the PCR amplification biases. In this context I propose to switch to other target enrichment techniques like DNA capture. Though this technique has proven successful in some studies, it still needs to evolve. In any case it could be quite interesting to develop methods for *in silico* capture like *in silico* PCR.

References

Acinas Silvia G., Sarma-Rupavtarm Ramahi, Klepac-Ceraj Vanja, and Polz Martin F. Pcr-induced sequence artifacts and bias: Insights from comparison of two 16s rrna clone libraries constructed from the same sample. *Applied and Environmental Microbiology*, 71 (12):8966–8969, 2005.

Agustí N., Shayler S. P., Harwood J. D., Vaughan I. P., Sunderland K. D., and Symondson W. O. C. Collembola as alternative prey sustaining spiders in arable ecosystems: prey detection within predators using molecular markers. *Molecular Ecology*, 12(12):3467–3475, 2003.

Aho Alfred V. and Corasick Margaret J. Efficient string matching: An aid to bibliographic search. *Communications of the ACM*, 6(18), 1975.

Altschul Stephen F., Madden Thomas L., Schäffer Alejandro A., Zhang Jinghui, Zhang Zheng, Miller Webb, and Lipman David J. Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic Acids Research*, 25(17):3389–3402, 1997.

Apostolico A. and Preparata F. P. Structural properties of the string statistics problem. *Journal of Computer and System Sciences*, 31:394–411, 1985.

Arnot David E., Roper Caroline, and Bayoumi Riad A. L. Digital codes from hypervariable tandemly repeated dna sequences in the plasmodium falciparum circumsporozoite gene can genetically barcode isolates. *Molecular and biochemical parasitology*, 61(1):15–24, 1993.

Ashelford Kevin E., Chuzhanova Nadia A., Fry John C., Jones Antonia J., and Weightman Andrew J. At least 1 in 20 16s rrna sequence records currently held in public repositories is estimated to contain substantial anomalies. *Applied and Environmental Microbiology*, 71(12):7724–7736, 2005.

Baer Jean-Loup and Lin Yi-Bing. Improving quicksort performance with a codeword data structure. *IEEE Transactions on Software Engineering*, 15(5):622–631, 1989.

Battiti Roberto and Protasi Marco. Reactive search, a history-based heuristic for max-sat. *ACM Journal of Experimental Algorithmics*, 2, 1997.

Beer Stafford. Application of modern heuristic methods. *Journal of The Operational Research Society*, 47:715–716, 1996.

Bekaert Michaël and Teeling Emma C. Uniprime: a workflow-based platform for improved universal primer design. *Nucleic Acids Research*, 36(10):e56+, 2008.

Bellemain Eva, Carlsen Tor, Brochmann Christian, Coissac Eric, Taberlet Pierre, and Kauserud Håvard. Its as an environmental dna barcode for fungi: an in silico approach reveals potential pcr biases. *BMC Microbiology*, 10:189, 2010.

Bentley D. R. Whole genome resequencing. *Current Opinion in Genetics and Development*, 16(6):545 – 552, 2006.

Blackwood Christopher B., Hudleston Deborah, Zak Donald R., and Buyer Jeffrey S. Interpreting ecological diversity indices applied to terminal restriction fragment length polymorphism data: Insights from simulated microbial communities. *Applied and Environmental Microbiology*, 73(16):5276–5283, 2007.

Blaxter Mark, Mann Jenna, Chapman Tom, Thomas Fran, Whitton Claire, Floyd Robin, and Eyualem-Abebe. Defining operational taxonomic units using DNA barcode data. *Philosophical Transactions Of The Royal Society B: Biological Sciences*, 360(1462):1935–1943, 2005.

Blum Christian and Roli Andrea. Metaheuristics in combinatorial optimization: Overview and conceptual comparison. *ACM computing surveys*, 35(3):268–308, 2003.

Boessenkool S., Epp L. S., Haile J., Bellemain E., Eposito A., and Coissac E. Arctic archives in the dirt: Paleoecological reconstruction of biodiversity from ancient dna preserved in permafrost soils. Oslo Science Conference, 2010.

Born Coen and Kerbosch Joep. Algorithm 457: finding all cliques of an undirected graph. *Communications of the ACM*, 16(9), 1973.

Briggs Adrian W., Stenzel Udo, Johnson Philip L. F., Green Richard E., Kelso Janet, Prüfer Kay, Meyer Matthias, Krause Johannes, Ronan Michael T., Lachmann Michael, and Pääbo Svante. Patterns of damage in genomic dna sequences from a neandertal. *Proceedings of the National Academy of Sciences of United States of America*, 104(37):14616–14621, 2007.

Briggs Adrian W., Good Jeffrey M., Green Richard E., Krause Johannes, Maricic Tomislav, Stenzel Udo, Lalueza-Fox Carles, Rudan Pavao, Brajković Dejana, Kućan Željko, Gušić Ivan, Schmitz Ralf, Doronichev Vladimir B., Golovanova Liubov V., de la Rasilla Marco, Fortea Javier, Rosas Antonio, and Pääbo Svante. Targeted retrieval and analysis of five neandertal mtdna genomes. *Science*, 325(5938):318–321, 2009.

Burbano Hernán A., Hodges Emily, Green Richard E., Briggs Adrian W., Krause Johannes, Meyer Matthias, Good Jeffrey M., Maricic Tomislav, Johnson Philip L. F., Xuan Zhenyu, Rooks Michelle, Bhattacharjee Arindam, Brizuela Leonardo, Albert Frank W., de la Rasilla Marco, Fortea Javier, Rosas Antonio, Lachmann Michael, Hannon Gregory J., and Pääbo Svante. Targeted investigation of the neandertal genome by array-based sequence capture. *Science*, 328(5979):723–725, 2010.

Burpo F. John. A critical review of pcr primer design algorithms and crosshybridization case study. *Biochemistry*, 218, 2001.

Campbell Anthony. Save those molecules! molecular biodiversity and life*. *Journal of Applied Ecology*, 40(2):193–203, 2003.

Church George M. Genomes for al. *Scientific American*, 294:46–54, 2006.

Clark Andrew G. and Whittam Thomas S. Sequencing errors and molecular evolutionary analysis. *Molecular Biology and Evolution*, 9(4):744–752, 1992.

Clarke L. A., Rebelo C. S., Gonçalves J., Boavida M. G., and Jordan P. Pcr amplification introduces errors into mononucleotide and dinucleotide repeat sequences. *Molecular Pathology*, 54(5):351–353, 2001.

Cline Janice, Braman Jeffery C., and Hogrefe Holly H. Pcr fidelity of pfu dna polymerase and other thermostable dna polymerases. *Nucleic Acids Research*, 24(18):3546–3551, 1996.

Cole J. R., Chai B., Marsh T. L., Farris R. J., Wang Q., Kulam S. A., Chandra S., McGarrell D. M., Schmidt T. M., Garrity G. M., and Tiedje J. M. The ribosomal database project (rdp-ii): previewing a new autoaligner that allows regular updates and the new prokaryotic taxonomy. *Nucleic Acids Research*, 31(1):442–443, 2003.

Connolly David T. An improved annealing scheme for the qap. *European Journal of Operational Research*, 46(1):93 – 100, 1990.

Cummings S. M., McMullan M., Joyce D. A., and van Oosterhout C. Solutions for pcr, cloning and sequencing errors in population genetic analysis. *Conservation Genetics*, 11: 1095–1097, 2010.

Darwin Charles. *The Origin of Species*. John Murray, Albemarble Street London, 1859.

de Hoon M.J.L., Imoto S., Nolan J., and Miyano S. Open source clustering software. *Bioinformatics*, 20(9):1453–1454, 2004.

Deagle Bruce, Eveson J. Paige, and Jarman Simon. Quantification of damage in DNA recovered from highly degraded samples - a case study on DNA in faeces. *Frontiers in Zoology*, 3(1):11+, 2006.

Dohm Juliane C., Lottaz Claudio, Borodina Tatiana, and Himmelbauer Heinz. Substantial biases in ultra-short read data sets from high-throughput dna sequencing. *Nucleic Acids Research*, 36(16):e105, 2008.

Dray Stéphane and Dufour Anne-Béatrice. The ade4 package: implementing the duality diagram for ecologists. *Journal of Statistical Software*, 22(4):1–20, 2007.

Duitama Jorge, Kumar Dipu M., Hemphill Edward, Khan Mazhar, Mandoiu Ion I., and Nelson Craig E. Primerhunter: a primer design tool for pcr-based virus subtype identification. *Nucleic Acids Research*, 37(8):2483–2492, 2009.

Ehrenfeucht A. and Haussler D. A new distance metric on strings computable in linear time. *Discrete Applied Mathematics*, 20:191–203, 1988.

Epp Laura S., Boessenkool Sanne, Bellemain Eva P., Haile James, Esposito Alfonso, Riaz Tiayyba, Erséus Christer, Gusarov Vladimir, Edwards Mary E., Johnsen Arild, Stenøien Hans, Hassel Kristian, Willerslev Eske, Taberlet Pierre, Coissac Eric, and Brochmann Christian. New environmental metabarcodes for analysing soil dna: potential for studying past and present ecosystems. 2011 .

Ereshefsky Marc. *The Poverty of the Linnaean Hierarchy: A Philosophical Study of Biological Taxonomy*. Cambridge University Press, 2001.

Ewing Brent and Green Phil. Base-calling of automated sequencer traces using phred. ii. error probabilities. genome res. *Genome Research*, 8:186–194, 1998.

Farach Martin. Optimal suffix tree construction with large alphabets. In *Proceedings of the 38th Annual Symposium on Foundations of Computer Science*, pages 137–148, Washington, DC, USA, 1997. IEEE Computer Society. ISBN 0-8186-8197-7.

Ficetola Gentile F., Coissac Eric, Zundel Stephanie, Riaz Tiayyba, Shehzad Wasim, Bessiere Julien, Taberlet Pierre, and Pompanon Francois. An in silico approach for the evaluation of dna barcodes. *BMC Genomics*, 11(1):434+, 2010.

Floyd Robin, Abebe Eyualem, Papert Artemis, and Blaxter Mark. Molecular barcodes for soil nematode identification. *Molecular Ecology*, 11(4):839–850, 2002.

Gadberry Michael D., Malcomber Simon T., Doust Andrew N., and Kellogg Elizabeth A. Primaclade-a flexible tool to find conserved pcr primers across multiple species. *Bioinformatics*, 21(7):1263–1264, 2005.

Gilbert M. Thomas P., Binladen Jonas, Miller Webb, Wiuf Carsten, Willerslev Eske, Poinar Hendrik, Carlson John E., Leebens-Mack James H., and Schuster Stephan C. Recharacterization of ancient dna miscoding lesions: insights in the era of sequencing-by-synthesis. *Nucleic Acids Research*, 35(1):1–10, 2007.

Glover Fred. Future paths for integer programming and links to artificial intelligence. *Computers and Operations Research*, 13(5):533–549, 1986.

Grant Verne. Incongruence between cladistic and taxonomic systems. *American Journal of Botany*, 90(9):1263–1270, 2003.

Gusfield Dan. *Algorithms on strings, trees, and sequences: computer science and computational biology*. Cambridge University Press, New York, USA, 1997. ISBN 0-521-58519-8.

Haas Brian J., Gevers Dirk, Earl Ashlee M., Feldgarden Mike, Ward Doyle V., Giannoukos Georgia, Ciulla Dawn, Tabbaa Diana, Highlander Sarah K., Sodergren Erica, Methé Barbara, DeSantis Todd Z., Consortium The Human Microbiome, Petrosino Joseph F., and Birren Rob Knightand Bruce W. Chimeric 16s rrna sequence formation and detection in sanger and 454-pyrosequenced pcr amplicons. *Genome Research*, 21(3):494–504, 2011.

Hajibabaei Mehrdad, Singer Gregory, Hebert Paul, and Hickey Donal. DNA barcoding: how it complements taxonomy, molecular phylogenetics and population genetics. *Trends in genetics*, 23(4):167–172, 2007.

Hajibabaei Mehrdad, hadi Shokralla, Zhou Xin, Baird Donald J., and Hebert Paul D.N. Environmental barcoding: A next generation sequencing approach to biodiversity monitoring. In *NABS 57th Annual Meeting*. North American Benthological Society, 2009.

Hall Neil. Advanced sequencing technologies and their wider impact in microbiology. *The Journal of Experimental Biology*, 210(9):1518–1525, 2007.

Hansen Anders J., Willerslev Eske, Wiuf Carsten, Mourier Tobias, and Arctander Peter. Statistical evidence for miscoding lesions in ancient dna templates. *Molecular Biology and Evolution*, 18(2):262–265, 2001.

Harris D. James. Can you bank on genbank? *Trends in Ecology and Evolution*, 18(7):317 – 319, 2003.

Hebert Paul D. N., Cywinska Alina, Ball Shelley L., and deWaard Jeremy R. Biological identifications through dna barcodes. *Proceedings. Biological Sciences*, 270(1512):313–321, 2003a.

Hebert Paul D. N., Ratnasingham Sujeevan, and deWaard Jeremy R. Barcoding animal life: cytochrome c oxidase subunit 1 divergences among closely related species. *Proceedings. Biological Sciences*, 270 Suppl 1:96–99, 2003b.

Herrera Aude, Héry Marina, Stach James E.M., Jaffré Tanguy, Normand Philippe, and Navarro Elisabeth. Species richness and phylogenetic diversity comparisons of soil microbial communities affected by nickel-mining and revegetation efforts in new caledonia. *European Journal of Soil Biology*, 43(2):130 – 139, 2007.

Hey Jody. The ancestor's tale a pilgrimage to the dawn of evolution. *The Journal of Clinical Investigation*, 115(7):1680–1680, 2005.

Hickerson Michael J., Meyer Christopher P., and Moritz Craig. Dna barcoding will often fail to discover new animal species over broad parameter space. *Systematic Biology*, 55 (5):729–739, 2006.

Hofreiter M., Poinar H. N., Paulding W. G. S, Bauer K., Martin P. S., Possnert G., and Paabo S. A molecular analysis of ground sloth diet through the last glaciation. *Molecular ecology*, 9(12):1975–1984, 2000.

Hofreiter Michael, Jaenicke Viviane, Serre David, Haeseler Arndt von, and Pääbo Svante. Dna sequences from multiple amplifications reveal artifacts induced by cytosine deamination in ancient dna. *Nucleic Acids Research*, 29(23):4793–4799, 2001.

Hohl Michael, Kurtz Stefan, and Ohlebusch Enno. Efficient multiple genome alignment. *Bioinformatics*, 18(suppl_1):312–320, 2002.

Hollingswortha Peter M., Forresta Laura L., Spouge John L., and others 49. A dna barcode for land plants. *Proceedings of the National Academy of Sciences of United Dates of America*, 106(31):12794–12797, 2009.

Hu Jianjun, Li Bin, and Kihara Daisuke. Limitations and potentials of current motif discovery algorithms. *Nucleic Acids Research*, 33(15):4899–4913, 2005.

Huber Julie A., Welch David B. Mark, Morrison Hilary G., Huse Susan M., Neal Phillip R., Butterfield David A., and Sogin Mitchell L. Microbial population structures in the deep marine biosphere. *Science*, 318(5847):97–100, 2007.

Huber Thomas, Faulkner Geoffrey, and Hugenholtz Philip. Bellerophon: a program to detect chimeric sequences in multiple sequence alignments. *Bioinformatics*, 20(14): 2317–2319, 2004.

Hudson Matthew E. Sequencing breakthroughs for genomic ecology and evolutionary biology. *Molecular Ecology Resources*, 8(1):3–17, 2008.

Huse Susan M., Welch David Mark M., Morrison Hilary G., and Sogin Mitchell L. Ironing out the wrinkles in the rare biosphere through improved otu clustering. *Environmental microbiology*, 12(7):1889–1898, 2010.

Jabado Omar J., Palacios Gustavo, Kapoor Vishal, Hui Jeffrey, Renwick Neil, Zhai Junhui, Briese Thomas, and Lipkin W. Ian. Greene scprimer: a rapid comprehensive tool for designing degenerate primers from multiple sequence alignments. *Nucleic Acids Research*, 34(22):6605–6611, 2006.

Janzen Daniel H, Hallwachs Winnie, Blandin Patrick, Burns John M, Cadiou Jean-Marie, Chacon Isidro, Dapkey Tanya, Deans Andrew R, Epstein Marc E, Espinoza Bernardo, Franclemont John G, Haber William A, Hajibabaei Mehrdad, Hall Jason P W, Hebert Paul D N, and others 31. Integration of dna barcoding into an ongoing inventory of complex tropical biodiversity. *Molecular Ecology Resources*, 9:1–26, 2009.

Jarman S. N., Gales N. J., Tierney M., Gill P. C., and Elliott N. G. A dna-based method for identification of krill species and its application to analysing the diet of marine vertebrate predators. *Molecular Ecology*, 11(12):2679–2690, 2002.

Johnson Philip L. F. and Slatkin Montgomery. Accounting for bias from sequencing error in population genetic estimates. *Molecular Biology and Evolution*, 25(1):199–206, 2008.

Kaartinen Riikka, Stone Graham N., Hearn Jack, Lohse Konard, and Roslin Tomas. Revealing secret liaisons: Dna barcoding changes our understanding of food webs. *Ecological Entomology*, 35(5):623–638, 2010.

Kaderali Lars and Schliep Alexander. Selecting signature oligonucleotides to identify organisms using dna arrays. *Bioinformatics*, 18(10):1340–1349, 2002.

Karp Richard M. Reducibility among combinatorial problems. In Miller R. E. and Thatcher J. W., editors, *Complexity of Computer Computations*, pages 85–103. Plenum Press, 1972.

Karp Richard M., Miller Raymond E., and Rosenberg Arnold L. Rapid identification of repeated patterns in strings, trees and arrays. In *STOC '72: Proceedings of the fourth annual ACM symposium on Theory of computing*, pages 125–136, New York, USA, 1972. ACM.

Kim Namshin and Lee Christopher. Qprimer: a quick web-based application for designing conserved pcr primers from multigenome alignments. *Bioinformatics*, 23(17):2331–2333, 2007.

Kirkpatrick S., Gelatt C. D., and Vecchi M. P. Optimization by simulated annealing. *Science*, 220(4598):671–680, 1983.

Knowlton Nancy. *Annual Review of Ecology and Systematics.*, volume 24, chapter Sibling species in the sea, pages 189–216. Ann. Rev. Ecol. Sys., 1993.

Ko Pang and Aluru Srinivas. Space efficient linear time construction of suffix arrays. In *Journal of Discrete Algorithms*, pages 200–210. Springer, 2003.

Kobayashi Norio, Tamura Koichiro, and Aotsuka Tadashi. Pcr error and molecular population genetics. *Biochemical Genetics*, 37(9-10):317–321, 1999.

Kocher T. D., Thomas W. K., Meyer A., Edwards S. V., Pääbo S., Villablanca F. X., and Wilson A. C. Dynamics of mitochondrial dna evolution in animals: amplification and sequencing with conserved primers. *Proceedings of the National Academy of Sciences of United Dates of America*, 86(16):6196–6200, 1989.

Komatsoulis GA and Waterman MS. A new computational method for detection of chimeric 16s rrna artifacts generated by pcr amplification from mixed bacterial populations. *Applied and Environmental Microbiology*, 63(6):2338–2346, 1997.

Korbel Jan O., Urban Alexander Eckehart, Affourtit Jason P., Godwin Brian, Grubert Fabian, Simons Jan Fredrik, Kim Philip M., Palejev Dean, Carriero Nicholas J., Du Lei, Taillon Bruce E., Chen Zhoutao, Tanzer Andrea, Saunders A. C. Eugenia, Chi Jianxiang, Yang Fengtang, Carter Nigel P., Hurles Matthew E., Weissman Sherman M., Harkins Timothy T., Gerstein Mark B., Egholm Michael, and Snyder Michael. Paired-end mapping reveals extensive structural variation in the human genome. *Science*, 318 (5849):420–426, 2007.

Kress W. John and Erickson David L. DNA barcodes: Genes, genomics, and bioinformatics. *Proceedings of the National Academy of Sciences of the United States of America*, 105(8):2761–2762, 2008.

Kress W. John, Wurdack Kenneth J., Zimmer Elizabeth A., and Weigt Lee A. Use of dna barcodes to identify flowering plants. *Proceedings of the National Academy of Sciences of the United States of America*, 102(23):8369–74, 2005.

Kuch Melanie, Rohland Nadin, Betancourt Julio L., Latorre Claudio, Steppan Scott, and Poinar Hendrik N. Molecular analysis of a 11 700-year-old rodent midden from the atacama desert, chile. *Molecular Ecology*, 11(5):913–924, 2002.

Kunin Victor and Hugenholtz Philip. Pyrotagger: A fast, accurate pipeline for analysis of rrna amplicon pyrosequence data. *The Open Journal*, 1:1, 2010.

Kunin Victor, Engelbrektson Anna, Ochman Howard, and Hugenholtz Philip. Wrinkles in the rare biosphere: pyrosequencing errors can lead to artificial inflation of diversity estimates. *Environmental Microbiology*, 12(1):118–123, 2010.

Kurtz Stefan and Schleiermacher Chris. Reputer: fast computation of maximal repeats in complete genomes. *Bioinformatics*, 15(5):426–427, 1999.

Kwok S., Kellogg D.E., McKinney N., Spasic D., Goda L., Levenson C., and Sninsk J.J. Effects of primer-template mismatches on the polymerase chain reaction: Human immunodeficiency virus type 1 model studies. *Nucleic Acids Research*, 18(4):999–1005, 1990.

Lahaye Renaud, van der Bank Michelle, Bogarin Diego, Warner Jorge, Pupulin Franco, Gigot Guillaume, Maurin Olivier, Duthoit Sylvie, Barraclough Timothy G., and Savolainen Vincent. DNA barcoding the floras of biodiversity hotspots. *Proceedings of the National Academy of Sciences of United Dates of America*, 105(8):2923–2928, 2008.

Landau Gad M. and Vishkin Uzi. Fast parallel and serial approximate string matching. *Journal of Algorithms*, 10(2):157–169, 1989.

Lawrence Charles E., Altschul Stephen F., Boguski Mark S., Liu Jun S., Neuwald Andrew F., and Wootton John C. Detecting subtle sequence signals: a gibbs sampling strategy for multiple alignment. *Science*, 262(5131):208–214, 1993.

Lee James Chun-I, Tsai Li-Chin, Yang Chung-Yu, and others 4. Dna profiling of shahtoosh. *ELECTROPHORESIS*, 27(17):3359–3362, 2006.

Lin Yawling, Jiang Tao, and mao Chao Kun. Efficient algorithms for locating the length-constrained heaviest segments with applications to biomolecular sequence analysis. *Journal of Computer and System Sciences*, 65:570–586, 2002.

Linnaeus Carolus. *Systema naturae*, volume 1. Holmiæ (Salvius), 1758.

Livingstone Craig D. and Barton Geoffrey J. Protein sequence alignments: a strategy for the hierarchical analysis of residue conservation. *Computer applications in the biosciences*, 9(6):745–756, 1993.

Lourenço Helena Ramalhinho, Martin Olivier, and Stutzle Thomas. A beginner's introduction to iterated local search. In *MIC 2001-Metaheuristics International Conference*, volume 1, pages 1–6, 2001.

Lund C. and Yannakakis M. On the hardness of approximating minimization problems. *Journal of the ACM (JACM)*, 41(5):960–981, 1994.

Mamanova Lira, Coffey Alison J, Scott Carol E, Kozarewa Iwanka, Turner Emily H, Kumar Akash, Howard Eleanor, Shendure Jay, and Turner Daniel J. Target-enrichment strategies for next- generation sequencing. *Nature Methods*, 7(2):111 – 118, 2010.

Manber Udi and Myers Gene. Suffix arrays: a new method for on-line string searches. In *SODA '90: Proceedings of the first annual ACM-SIAM symposium on Discrete algorithms*, pages 319–327, Philadelphia, PA, USA, 1990. Society for Industrial and Applied Mathematics.

Margulies Marcel, Egholm Michael, Altman William E., Attiya Said, Bader Joel S., Bemben Lisa A., Berka Jan, Braverman Michael S., and *et al.* Yi-Ju Chen. Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, 437(7057):376–380, 2005.

Martinez Hugo M. An efficient method for finding repeats in molecular sequences. *Nucleic Acids Research*, 11(13):4629–4634., 1983.

Mayr Ernst. *Systematics and the Origin of Species*. Columbia University Press, New York, 1942.

Mayr Ernst and Bock Walter. Classifications and other ordering systems. *Journal of Zoological Systematics and Evolutionary Research*, 40(4):169–194, 2002.

McCreight Edward M. A space-economical suffix tree construction algorithm. *Journal of the ACM*, 23(2):262–272, 1976.

Min Xiang Jia and Hickey Donal A. Assessing the effect of varying sequence length on dna barcoding of fungi. *Molecular Ecology Notes*, 7(3):1–9, 2009.

Morgan Martin, Anders Simon, Lawrence Michael, Aboyoun Patrick, Pagès Hervé, and Gentleman Robert. Shortread: a bioconductor package for input, quality assessment and exploration of high-throughput sequence data. *Bioinformatics*, 25(19):2607–2608, 2009.

Nekrutenko Anton and Li WenHsiung. Assessment of compositional heterogeneity within and between eukaryotic genomes. *Genome Research*, 10(12):1986–1995, 2000.

Niemelä Jari. Biodiversity monitoring for decision-making. *Annales Zoologici Fennici*, 37: 307–317, 2000.

Notredame Cédric, Higgins Desmond G., and Heringa Jaap. T-coffee: a novel method for fast and accurate multiple sequence alignment. *Journal of Molecular Biology*, 302(1):205 – 217, 2000.

Nowicki Eugeniusz and Smutnicki Czeslaw. A fast tabu search algorithm for the permutation flow-shop problem. *European Journal of Operational Research*, 91(1):160 – 175, 1996.

Osman I. H. and Kelly J. P., editors. *Meta-Heuristics: Theory and Applications*. Kluwer academic publishers, 1996.

Osman Ibrahim and Laporte Gilbert. Metaheuristics: A bibliography. *Annals of Operations Research*, 63:511–623, 1996.

Pääbo Svante, Irwin David M., and Wilson Allan C. Dna damage promotes jumping between templates duringenzymaticamplification. *The Journal of Biological Chemistry*, 265(8):4718–4721, 1990.

Pääbo Svante, Poinar Hendrik, Serre David, Jaenicke-Després Viviane, Hebler Juliane, Rohland Nadin, Kuch Melanie, Krause Johannes, Vigilant Linda, and Hofreiter Michael. Genetic analyses from ancient dna. *Annual Review of Genetics*, 38(1):645–679, 2004.

Palumbi Stephen R. *Nucleic acids II*, chapter the polymerase chain reaction, pages 205–247. Molecular Systematics. Sinauer & Associates Inc, 1996.

Passmore A, Jarman SN, Swadling KM, Kawaguchi SR, McMinn A, and Nicol S. DNA as a dietary biomarker in antarctic krill, euphausia superba. *Marine Biotechnology*, 8(6): 686–696, 2006.

Pompanon François, Coissac Eric, and Taberlet Pierre. Metabarcoding, une nouvelle façon d'analyser la biodiversité. *Biofutur*, (319):30–32, 2011.

Quince Christopher, Lanzén Anders, Curtis Thomas P, Davenport Russell J, Hall Neil, Head Ian M, Read L Fiona, and Sloan William T. Accurate determination of microbial diversity from 454 pyrosequencing data. *Nature Methods*, 6(9):639–641, 2009.

Quince Christopher, Lanzen Anders, Davenport Russell, and Turnbaugh Peter. Removing noise from pyrosequenced amplicons. *BMC Bioinformatics*, 12(1):38, 2011.

Ratnasingham Sujeevan and Hebert Paul D. N. bold: The barcode of life data system (http://www.barcodinglife.org). *Molecular Ecology Notes*, 7(3):355–364, 2007.

Ray John. Historia plantarum species hactenus editas aliasque insuper multas noviter inventas & descriptas complectens... 1, 1686.

Ray John. Historia plantarum species hactenus editas aliasque insuper multas noviter inventas & descriptas complectens... 2, 1688.

Ray John. Historia plantarum species hactenus editas aliasque insuper multas noviter inventas & descriptas complectens... 3, 1704.

Reeder Jens and Knight Rob. Rapid denoising of pyrosequencing amplicon data: exploiting the rank-abundance distribution. *Nature Methods*, 7(9):668–669, 2010.

Reeves Colin R. *Modern heuristic techniques for combinatorial problems.* McGraw-Hill, London, 1995.

Rice Peter, Longden Ian, and Bleasby Alan. Emboss: The european molecular biology open software suite. *Trends in Genetics*, 16(6):276–277, 2000.

Roura Xavier Domingo, Marmi Josep, Ferrando Aïnhoa, López-Giráldeza Francesc, Macdonalde David W., and Jansmanf Hugh A.H. Badger hair in shaving brushes comes from protected eurasian badgers. *Biological Conservation*, 128(3):425–430, 2006.

Rozen Steve and Skaletsky Helen. Primer3 on the www for general users and for biologist programmers. *Methods in Molecular Biology*, 132:365–386, 2000.

Rubinoff Daniel, Cameron Stephen, and Will Kipling. Are plant dna barcodes a search for the holy grail? *Trends in Ecology and Evolution*, 21(1):1–2, 2006.

Saiki R K, Scharf S, Faloona F, Mullis K B, Horn G T, Erlich H A, and Arnheim N. Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. *Science*, 230(4732):1350–4, Dec 1985.

Saitou Naruya and Nei Masatoshi. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, 4(4):406–425, 1987.

Sanger F., Nicklen S., and Coulson A. R. Dna sequencing with chain-terminating inhibitors. *Proceedings of The National Academy of Sciences of United States of America*, 74(12):5463–5467, 1977.

SantaLucia John and Hicks Donald. The thermodynamics of dna structural motifs. *Annual Review of Biophysics and Biomolecular Structure*, 33:415–440, 2004.

Schneider Thomas D., Stormo Gary D., Gold Larry, and Ehrenfeucht Andrzej. Information content of binding sites on nucleotide sequences. *Journal of Molecular Biology*, 188(3): 415–431, 1986.

Schuster Stephan C. Next-generation sequencing transforms today's biology. *Nature Methods*, 5(1):16–18, 2008.

Shehzad Wasim, Riaz Tiayyba, Nawaz Muhammad Ali, Miquel Christian, Poillot Carole, Shah Safdar Ali, Pompanon François, Coissac Eric, and Taberlet Pierre. A universal approach for carnivore diet analysis based next generation sequencing: application to the leopard cat (prionailurus bengalensis) in pakistan.

Shendure Jay and Ji Hanlee. Next-generation dna sequencing. *Nature Biotechnology*, 26 (10):1135–1145, 2008.

Simpson Edward Hugh. Measurement of diversity. *Nature*, 163(4148):688–688, 1949.

Smith M. Alex, Woodley Norman E., Janzen Daniel H., Hallwachs Winnie, and Hebert Paul D. N. Dna barcodes reveal cryptic host-specificity within the presumed polyphagous members of a genus of parasitoid flies (diptera: Tachinidae). *Proceedings of the National Academy of Sciences of United Dates of America*, 103(10):3657–3662, 2006.

Sneath Peter Henry and Robert Sokal. *Numerical taxonomy the principles and practice of numerical classification*. W.H.Freeman, San Francisco, 1973.

Sogin Mitchell L., Morrison Hilary G., Huber Julie A., Welch David Mark, Huse Susan M., Neal Phillip R., Arrieta Jesus M., and Herndl Gerhard J. Microbial diversity in the deep sea and the underexplored "rare biosphere". *Proceedings of the National Academy of Sciences of the United States of America*, 103(32):12115–12120, 2006.

Soldano Henri, Viari Alain, and Champesm Marc. Searching for flexible repeated patterns using a non transitive similarity relation. *Pattern Recognition Letters*, 16:233–245, 1995.

Stojanovic Nikola, Florea Liliana, Riemer Cathy, Gumucio Deborah, Slightom Jerry, Goodman Morris, Miller Webb, and Hardison Ross. Comparison of five methods for finding conserved sequences in multiple alignments of gene regulatory regions. *Nucleic Acids Research*, 27(19):3899–3910, 1999.

Stützle Thomas. *Local Search Algorithms for Combinatorial Problems - Analysis, Algorithms and New Applications*. PhD thesis, 1999.

Taberlet Pierre, Coissac Eric, Pompanon François, Gielly Ludovic, Miquel Christian, Valentini Alice, Vermat Thierry, Corthier Gérard, Brochmann Christian, and Willerslev Eske. Power and limitations of the chloroplast trnl (uaa) intron for plant dna barcoding. *Nucleic Acids Research*, 35(3):e14, 2007.

Thompson Julie D., Gibson Toby J., Plewniak Frédéric, Jeanmougin François, and Higgins Desmond G. The CLUSTALX windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Research*, 25(24): 4876–4882, 1997.

Ukkonen Esko. On-line construction of suffix trees. *Algorithmica*, 14:249–260, 1995.

Vaessens R. J. M., Aarts E. H. L., and Lenstra J. K. Job shop scheduling by local search. *INFORMS Journal on Computing*, 8(3):302–317, 1996.

Valentini Alice, Pompanon François, and Taberlet Pierre. Dna barcoding for ecologists. *Trends in Ecology and Evolution*, 24:110–117, 2009.

van Laarhoven Peter J. M., Aarts Emile H. L., and Lenstra Jan K. Job shop scheduling by simulated annealing. *Opearations Research*, 40(1):113–125, 1992.

Černý V. Thermodynamical approach to the traveling salesman problem: An efficient simulation algorithm. *Journal of Optimization Theory and Applications*, 45(1):41–51, 1985.

Vernooy Ronnie, Haribabu Ejnavarzala, Muller Manuel Ruiz, Vogel Joseph Henry, Hebert Paul D. N., Schindel David E., Shimura Junko, and Singer Gregory A. C. Barcoding life to conserve biological diversity: Beyond the taxonomic imperative. *PLoS Biology*, 8(7): e1000417, 2010.

Vicente Gomez Alvarez, M King Gary, and Klaus Nusslein. Comparative bacterial diversity in recent hawaiian volcanic deposits of different ages. *FEMS Microbiology Ecology*, pages 60–73, 2007.

Vidal Rene V. V., editor. *Applied Simulated Annealing (Lecture Notes in Economics and Mathematical Systems)*. Springer-Verlag, Berlin, 1993.

Voss Stefan, Osman Ibrahim H., and Roucairol Catherine. *Meta-Heuristics: Advances and Trends in Local Search Paradigms for Optimization*. Kluwer Academic Publishers, Norwell, MA, USA, 1999. ISBN 0792383699.

Wang Gary P., Sherrill-Mix Scott A., Chang Kyong-Mi, Quince Chris, and Bushman Frederic D. Hepatitis c virus transmission bottlenecks analyzed by deep sequencing. *Journal Of Virology*, 84(12):6218–6228, 2010.

Wang Grace C. Y. and Wang Yue. The frequency of chimeric molecules as a consequence of pcr co-amplification of 16s rrna genes from different bacterial species. *Microbiology*, 142(1107-1114), 1996.

Weiner Peter. Linear pattern matching algorithms. In *Proceedings of the 14th Annual Symposium on Switching and Automata Theory*, pages 1–11, Washington, DC, USA, 1973. IEEE Computer Society.

Wheeler David L., Barrett Tanya, Benson Dennis A., and other authors 29. Database resources of the national center for biotechnology information. *Nucleic Acids Research*, 34(suppl 1):D173–D180, 2006.

Whittaker Robert J. Meta-analyses and mega-mistakes: calling time on meta-analysis of the species richness-productivity relationship. *Ecology*, 91(9):2522–2533, 2010.

Wilcox Bruce. In situ conservation of genetic resources: determinants of minimum area requirements. In *National Parks, Conservation and Development, Proceedings of the World Congress on National Parks*, pages 18–30. Smithsonian Institution Press, 1984.

Willerslev Eske, Hansen Anders J, Binladen Jonas, Brand Tina B, Gilbert M Thomas P, Shapiro Beth, Bunce Michael, Wiuf Carsten, Gilichinsky David A, and Cooper Alan. Diverse plant and animal genetic records from holocene and pleistocene sediments. *Science*, 300(5620):791–795, 2003.

Willerslev Eske, Cappellini Enrico, Boomsma Wouter, and others 27. Ancient biomolecules from deep ice cores reveal a forested southern greenland. *Science*, 317(5834):111–114, 2007.

Willi Henning. *Phylogenetic systematics*. Urbana: University of Illinois Press, 3rd edition, 1979.

Wintzingerode Friedrich V., Goëbel Ulf B., and Stackebrandt Erko. Determination of microbial diversity in environmental samples: pitfalls of pcr-based rrna analysis. *FEMS Microbiology Reviews*, 21(3):213–229, 1997.

Woese. Carl R., Kandler Otto, and Wheelis Mark L. Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. *Proceedings of the National Academy of Sciences of the United States of America*, 87(12):4576–4579, 1990.

Zhang Junbin, Huang Liangmin, and Huo Heqiang. Larval identification of lutjanus bloch in nansha coral reefs by aflp molecular method. *Journal of Experimental Marine Biology and Ecology*, 298(1):3 – 20, 2004.

Zinger Lucie, Gury Jérôme, Giraud Frédéric, Krivobok Serge, Gielly Ludovic, Taberlet Pierre, and Geremia Roberto A. Improvements of polymerase chain reaction and capillary electrophoresis single-strand conformation polymorphism methods in microbial ecology: Toward a high-throughput method for microbial diversity studies in soil. *Microbial Ecology*, 54(2):203–216, 2007.

Zinger Lucie, Gury Jérôme, Alibeu Olivier, Rioux Delphine, Gielly Ludovic, Sage Lucile, Pompanon François, and Geremia Roberto A. Ce-sscp and ce-fla, simple and high-throughput alternatives for fungal diversity studies. *Journal of Microbiological Methods*, 72(1):42 – 53, 2008.

Annex A
Article 3

Molecular Ecology (2011) doi: 10.1111/j.1365-294X.2011.05424.x

Carnivore diet analysis based on next-generation sequencing: application to the leopard cat (*Prionailurus bengalensis*) in Pakistan

WASIM SHEHZAD,* TIAYYBA RIAZ,* MUHAMMAD A. NAWAZ,*† CHRISTIAN MIQUEL,* CAROLE POILLOT,* SAFDAR A. SHAH,‡ FRANÇOIS POMPANON,* ERIC COISSAC* and PIERRE TABERLET*

**Laboratoire d'Ecologie Alpine, CNRS-UMR 5553, Université Joseph Fourier, BP 53, F-38041 Grenoble Cedex 9, France, †Snow Leopard Trust (Pakistan Program), 17-Service Road North, I-8/3, Islamabad, Pakistan, ‡Wildlife Department, Khyber* 1 *Pakhtunkhwa, Pakistan*

Abstract

Diet analysis is a prerequisite to fully understand the biology of a species and the functioning of ecosystems. For carnivores, traditional diet analyses mostly rely upon the morphological identification of undigested remains in the faeces. Here, we developed a methodology for carnivore diet analyses based on the next-generation sequencing. We applied this approach to the analysis of the vertebrate component of leopard cat diet in two ecologically distinct regions in northern Pakistan. Despite being a relatively common species with a wide distribution in Asia, little is known about this elusive predator. We analysed a total of 38 leopard cat faeces. After a classical DNA extraction, the DNA extracts were amplified using primers for vertebrates targeting about 100 bp of the mitochondrial 12S rRNA gene, with and without a blocking oligonucleotide specific to the predator sequence. The amplification products were then sequenced on a next-generation sequencer. We identified a total of 18 prey taxa, including eight mammals, eight birds, one amphibian and one fish. In general, our results confirmed that the leopard cat has a very eclectic diet and feeds mainly on rodents and particularly on the Muridae family. The DNA-based approach we propose here represents a valuable complement to current conventional methods. It can be applied to other carnivore species with only a slight adjustment relating to the design of the blocking oligonucleotide. It is robust and simple to implement and allows the possibility of very large-scale analyses.

Keywords: blocking oligonucleotide, DNA metabarcoding, mitochondrial DNA, ribosomal DNA, species identification

Received 8 June 2011; revision received 16 November 2011; accepted 24 November 2011

Introduction

The nature of trophic interactions is a fundamental question in ecology and has commanded the attention of biologists for decades. Dietary behavioural studies provide key data for understanding animal ecology, evolution and conservation (Symondson 2002; Krahn *et al.* 2007). Wild felids are among the keystone predators and have significant effects on ecosystem functioning, despite their relatively low biomass (Mills *et al.* 1993; Power *et al.* 1996). The modal mass concept (Macdonald *et al.* 2010) proposes that each felid species focuses on large-as-possible prey to maximize their intake relative to their energy expenditure for each catch, provided that such prey can be safely killed.

Owing to their elusive behaviour, scientific knowledge of South Asian wild cats is limited (Nowell & Jackson 1996). The leopard cat (*Prionailurus bengalensis*) is a small felid (weight 1.7–7.1 kg; Sunquist & Sunquist 2009), with a wide range in Asia (8.66×10^6 km^2; Nowell & Jackson 1996). Beginning in Pakistan and

Correspondence: Pierre Taberlet, Fax: +33(0)4 76 51 42 79; E-mail: pierre.taberlet@ujf-grenoble.fr

parts of Afghanistan in the west, the leopard cat occurs throughout Southeast Asia, including the islands of Sumatra, Borneo, and Taiwan. It extends into China, Korea, Japan and the Far East of Russia. (Macdonald *et al.* 2010). The leopard cat's flexible habitat selection and prey choices favour its distribution throughout the range (Watanabe 2009; Mukherjee *et al.* 2010). It is found in very diverse environments, from semideserts to tropical forests, woodlands to pine forests and scrubland to agriculture land (Sunquist & Sunquist 2002). It prefers to live in habitats near sources of water and can be found in the close proximity to human population (Scott *et al.* 2004).

The population status of the leopard cat is not uniform throughout its range. The cat is relatively secure in China (Lau *et al.* 2010) and in India (Nowell & Jackson 1996), endangered in Korea (Rho 2009) and most endangered in Japan (Mitani *et al.* 2009). In Pakistan, this species is categorized by the IUCN as ''data deficient'' as no information exists about the extent of its occurrence, nor its occupancy, population and habitat (Sheikh & Molur 2004). Major threats to the species include hunting, habitat loss and fragmentation because of the human population expansion in addition to competition for prey with other sympatric carnivores (Izawa & Doi 1991). Commercial exploitation for the fur trade is a significant threat throughout its range (Sheikh & Molur 2004); in China, the annual pelt harvest was estimated at to be 400 000 animals in mid-1980s (Nowell & Jackson 1996).

Despite being a relatively common species with a wide distribution, comparatively little information is available about the diet of the leopard cat in general, and no information at all specific to Pakistan, where this predator is rare. Faeces analysis by hair mounting and bone examination is used extensively and can provide information about the diet (e.g. Oli *et al.* 1994; Gaines 2001; Bagchi & Mishra 2006; Lovari *et al.* 2009). Muridae (mainly *Rattus* spp. and *Mus* spp.) seem to represent the main prey items throughout the leopard cat distribution range, supplemented by a wide variety of other prey including small mammals such as shrews and ground squirrels, birds, reptiles, frogs and fish (Tatara & Doi 1994; Grassman *et al.* 2005; Austin *et al.* 2007; Rajaratnam *et al.* 2007; Watanabe 2009; Fernandez *et al.* 2011). [2]

Molecular analysis of faeces (Höss *et al.* 1992; Kohn & Wayne 1997) provides an alternative noninvasive approach to study animal diet, but prey DNA in faeces is often highly degraded, preventing the amplification of long fragments (Zaidi *et al.* 1999; Jarman *et al.* 2002). Until 2009, most of the molecular-based studies to analyse diet were carried out using traditional sequencing approaches (e.g. Deagle *et al.* 2005, 2007; Bradley *et al.* 2007). These methods require cloning PCR products and subsequent Sanger sequencing of these clones by capillary electrophoresis. However, this approach is both time-consuming and expensive (Pegard *et al.* 2009).

Next-generation sequencing is revolutionizing diet analysis based on faeces (Valentini *et al.* 2009b), because sequence data from very large numbers of individual DNA molecules in a complex mixture can be studied without the need for cloning. Valentini *et al.* (2009a) have presented a universal approach for the diet analysis of herbivores. The methodology consists of extracting DNA from faeces to amplify it using the universal primers *g* and *h*, which amplify the short P6 loop of the chloroplast *trn*L (UAA) intron (Taberlet *et al.* 2007), and in sequencing the PCR products using a next-generation sequencer.

While such an approach has been successfully implemented for herbivores, the analysis of carnivore diet presents a real challenge when using primers for mammals or vertebrates, as predator DNA can be simultaneously amplified with prey DNA (Jarman *et al.* 2006; Deagle & Tollit 2007). Furthermore, prey fragments [3] might be rare in the DNA extract from faeces, and consequently be prone to being missed during the early stages of PCR, resulting in a PCR product almost exclusively containing the dominant sequences of predators (Jarman *et al.* 2004, 2006; Green & Minz 2005). Various methods have been proposed to avoid amplifying predator DNA. Species-specific or group-specific primers have been specially designed to avoid priming on predator DNA and to specifically amplify the target prey species (Vestheim *et al.* 2005; Deagle *et al.* 2006; King *et al.* 2010). This is not a convenient strategy if the prey are taxonomically diverse, which makes the design of suitable primers difficult (Vestheim & Jarman 2008). Another strategy involves cutting predator sequences with restriction enzymes before and/or during and/or after PCR amplification (Blankenship & Yayanos 2005; Green & Minz 2005; Dunshea 2009). However, these approaches can only be implemented with *a priori* knowledge of the potential prey.

The ideal system for studying carnivore diet using DNA in faeces lies in combining, in the same PCR, primers for vertebrates and a blocking oligonucleotide with a 3-carbon spacer (C3-spacer) on the 3′ end that specifically reduces the amplification of the predator DNA. Such a blocking oligonucleotide must be specifically designed to target predator DNA and thus bind preferentially with predator sequences, limiting their amplification. This concept has been effectively used in the field of clinical chemistry (Kageyama *et al.* 2008; Wang *et al.* 2008; Li *et al.* 2009) and in environmental microbiology (Liles *et al.* 2003). However, the application of blocking oligonucleotide in trophic studies is relatively recent. Vestheim & Jarman (2008) first used a blocking oligonucleotide to assess the diet of Antarctic krill. More recently, Deagle *et al.* (2009, 2010) investigated the diet of Australian fur seals (*Arctocephalus pusillus*) and penguins (*Eudyptula minor*) by combining a blocking oligonucleotide approach with 454 GS-FLX pyrosequencing technologies.

The main aim of this study was to analyse the leopard cat diet in two distinct environments in Pakistan by developing a method that would give the vertebrate diet profile of a carnivore without any *a priori* information about the prey species. This method is based on the use of recently designed primers for vertebrates (Riaz *et al.* 2011) together with a blocking oligonucleotide specific to the leopard cat and employing a high-throughput next-generation sequencer. However, such an approach cannot detect the cases of infanticide and possible cannibalism that have been documented in Felidae (e.g. Natoli 1990).

Materials and methods

General strategy for diet analysis of the leopard cat

Figure 1 outlines the general strategy we followed for the diet analysis of the leopard cat. After the faeces collection and DNA extraction, the samples were confirmed to be those of leopard cat by using leopard cat–specific primers. Selected samples were amplified in two series of experiments, one with primers for vertebrates and the other with the same primers plus a blocking oligonucleotide specific to the leopard cat. These PCR products were subsequently sequenced using the Illumina sequencing platform GA IIx. The amplified sequences of prey taxa were identified by comparison with reference databases (GenBank/EMBL/DDBJ), taking into account prey availability according to their geographic distributions.

Sample collection and preservation

Putative felid faeces were collected in two areas: Ayubia National Park (ANP) and Chitral Gol National Park (CGNP). Both national parks are located in the Khyber Pakhtunkhwa province and represent two extremities of the leopard cat range in Pakistan (Fig. 2). These national parks have disparate environments. The ANP is comprised of moist temperate forests, subalpine meadows and subtropical pine forests. Mean temperatures range between 4.2 °C in January to 26 °C in July. The altitudinal variation ranges from 1050 to 3027 m, and the mean annual rainfall is between 1065 and 1424 mm. It has ~200 species of birds, 31 species of

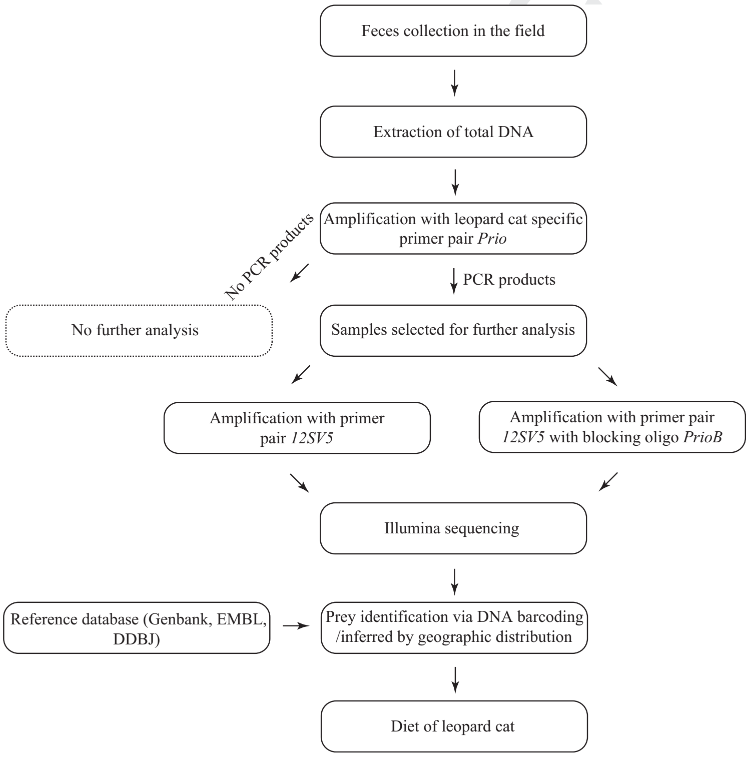

Feces collection in the field
Extraction of total DNA
Amplification with leopard cat specific primer pair *Prio*
No PCR products
PCR products
No further analysis
Samples selected for further analysis
Amplification with primer pair *12SV5*
Amplification with primer pair *12SV5* with blocking oligo *PrioB*
Illumina sequencing
Reference database (Genbank, EMBL, DDBJ)
Prey identification via DNA barcoding /inferred by geographic distribution
Diet of leopard cat

Fig. 1 Flowchart diagram showing the various steps involved in the diet analysis of the leopard cat. The samples in the dotted box were discarded from further experimentation.

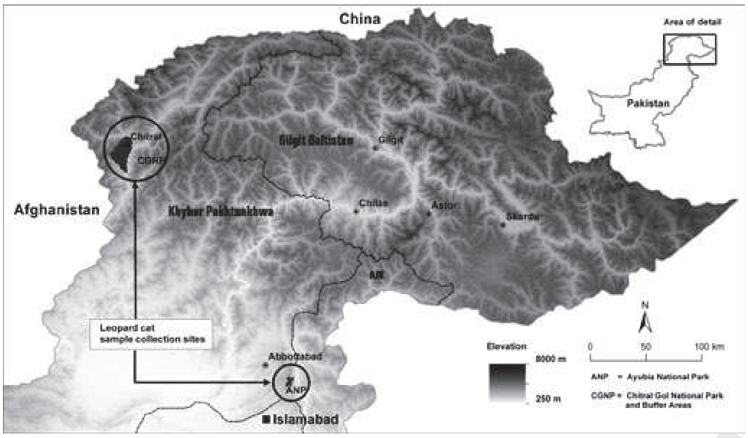


Fig. 2 Sampling locations of leopard cat faeces in northern Pakistan.

mammals, 16 species of reptiles and three species of amphibians (Farooque 2007).

The CGNP generally falls into a subtropical zone with vegetation classified as dry temperate forests. Forests of the park are growing under the extremes of climatic and edaphic factors, and tree canopy is rarely closed. Mean temperature of the valley ranges between 1 °C in January to 24 °C in July, and average annual rainfall varies between 450 and 600 mm. The park supports 29 mammals, 127 birds and nine reptiles (GoN-WFP & IUCN 1996; Mirza 2003).

We collected 114 faecal samples from ANP and 67 from the CGNP. The samples were preserved first in 90% ethanol and then shifted into silica gel for transportation to LECA (Laboratoire d'Ecologie Alpine), Université Joseph Fourier, Grenoble, France, for diet analysis.

DNA extraction

All extractions were performed in a room dedicated to degrade DNA extractions. Total DNA was extracted from about 15 mg of faeces using the DNeasy Blood and Tissue Kit (QIAgen GmbH). Each 15 mg faecal sample was incubated for at least 3 h at 55 °C with a lysis buffer (Tris–HCl 0.1 M, EDTA 0.1 M, NaCl 0.01 M and *N*-lauroyl sarcosine 1% with pH 7.5–8), before following the manufacturer's instructions. The DNA extracts were recovered in a total volume of 250 μL. Blank extractions without samples were systematically performed to monitor possible contaminations.

Selection/designing of primer pairs for the leopard cat diet study

Identification of faecal samples as leopard cat. We used the leopard cat–specific primer pair *PrioF/PrioR*, amplifying a 54-bp fragment (without primers) of the mitochondrial 12S gene (Table 1). The specificity of this primer pair was validated both by empirical experiments (Ficetola *et al.* 2010) and by the program *ecoPCR* (Bellemain *et al.* 2010; Ficetola *et al.* 2010), with parameters to prevent mismatches on the two last nucleotides of each primer, and designed to tolerate a maximum of three mismatches on the remaining part of the primers. The goal of such an experimental validation was to distinguish leopard cat faeces from those from the two other felid species potentially occurring in the study areas, i.e. the common leopard (*Panthera pardus*) in ANP and the snow leopard (*Panthera uncia*) in CGNP. The primary identification of samples was carried out on the basis of the presence of a PCR product of the suit-

Table 1 Sequences of the primer pairs used in the study. The length of amplified fragments (excluding primers) with *Prio* & *12SV5* was 54 and ~100 bp, respectively

Name	Primer sequence (5–3′)	References
PrioF	CCTAAACTTAGATAGTTAATTTT	Ficetola *et al.* (2010)
PrioR	GGATGTAAAGCACCGCC	Ficetola *et al.* (2010)
12SV5F	TAGAACAGGCTCCTCTAG	Riaz *et al.* (2011)
12SV5R	TTAGATACCCCACTATGC	Riaz *et al.* (2011)
PrioB	CTATGCTTAGCCCTAAACTTAGATAGTTAATTTTAACAAAACTATC-C3	This study

able length as revealed by electrophoresis on a 2% agarose gel. The samples successfully amplified using *PrioF/PrioR* were selected for further analyses.

The PCRs were carried out in a total volume of 20 μL with 8 mM Tris–HCl (pH 8.3), 40 mM KCl, 2 mM $MgCl_2$, 0.2 mM of each dNTP, 0.2 μM of each primer, BSA (5 μg), 0.5 U of AmpliTaq Gold® DNA polymerase (Applied Biosystems) using 2 μL of DNA extract as a template. The PCR conditions were set as an initial 10-min denaturation step at 95 °C to activate the polymerase, followed by 45 cycles of denaturation at 95 °C for 30 s and annealing at 53 °C for 30 s, without elongation steps as the amplified fragment was very short.

Blocking oligonucleotide specific to leopard cat sequences. The *PrioB* (Table 1) blocking oligonucleotide specific to leopard cat sequences was designed as suggested by Vestheim & Jarman (2008). This blocking oligonucleotide was used to limit the amplification of leopard cat sequences when using the primers targeting all vertebrates. Table 2 presents a sequence alignment of *PrioB* with the main groups of vertebrates. This blocking oligonucleotide might also slightly block the amplification of other felid species, but will not prevent the amplification of other vertebrate groups.

Primer pair for vertebrates. We used the primer pair *12SV5F/12SV5R* designed by the *ecoPrimers* program (Riaz *et al.* 2011). *ecoPrimers* scans whole genomes to find new barcode markers and their associated primers, by optimizing two quality indices measuring the taxonomical coverage and the discrimination power to select the most efficient markers, according to specific experimental constraints such as marker length or targeted taxa. This primer pair for vertebrates represents the best choice found by *ecoPrimers* among short barcodes, as derived from the available vertebrate whole mitochondrial genomes currently available. It amplifies a ~100-bp fragment of the V5 loop of the mitochondrial 12S gene, with the ability to amplify short DNA fragments such as those recovered from faeces, and has a high taxonomic resolution, despite its short size. Using the *ecoPCR* program (Bellemain *et al.* 2010; Ficetola *et al.* 2010), and based on the release 103 of the EMBL database, this fragment unambiguously identifies 77% of the species and 89% of the genera as recorded by this EMBL release (Riaz *et al.* 2011).

DNA amplification for diet analysis

All DNA amplifications were carried out in a final volume of 25 μL, using 2 μL of DNA extract as template. The amplification mixture contained 1 U of AmpliTaq Gold® DNA Polymerase (Applied Biosystems), 10 mM Tris–HCl, 50 mM KCl, 2 mM of $MgCl_2$, 0.2 mM of each dNTP, 0.1 μM of each primer (*12SV5F/12SV5R*) and 5 μg of bovine serum albumin (BSA; Roche Diagnostic). The PCR mixture was denatured at 95 °C for 10 min, followed by 45 cycles of 30 s at 95 °C and 30 s at 60 °C; as the target sequences are ~100 bp long, the elongation step was removed to reduce the +A artefact (Brownstein *et al.* 1996; Magnuson *et al.* 1996) that might decrease the efficiency of the first step of the sequencing process (blunt-end ligation). Using the aforementioned conditions, the DNA extracts were amplified twice, first with *12SV5F/12SV5R* (0.1 μM each) and second with *12SV5F/12SV5R/PrioB* (0.1 μM for *12SV5F* and *12SV5R*, 2 μM for *PrioB*). These primer concentrations have been chosen after a series of test experiments, with various concentrations of *PrioB* (data not shown).

Table 2 Sequence alignment showing the specificity of the *PrioB* blocking oligonucleotide. The first six nucleotides of the *PrioB* blocking oligonucleotide overlap with the *12SV5R* amplification primer. This sequence alignment contains two other Felidae species (*Felis catus* and *Panthera tigris*), another carnivore species from the Ursidae family (*Ursus arctos*), two rodents (*Rattus rattus* and *Microtus kikuchii*), one insectivore (*Crocidura russula*), one bird (*Gallus gallus*), one amphibian (*Rana nigromaculata*) and one fish (*Cyprinus carpio*)

Accession number	Species name	Sequences (5′–3′)
PrioB blocking oligonucleotide		CTATGCTTAGCCCTAAACTTAGATAGTTAATTTTAACAAAACTATC
HM185183	*Prionailurus bengalensis*	....
NC_001700	*F. catus*	............................CCC.A............
JF357967	*P. tigris*	..................C...........CCCA............
NC_003427	*U. arctos*	............T.....A..A...A..T...AA.CA...TTAT..
NC_012374	*R. rattus*	..................C.TA...A...CA.C..CA....TAT.T
NC_003041	*M. kikuchii*	..................C.TAG..A..TTAAAAC.A...TA.T.G
NC_006893	*C. russula*	...................A.A.C.A.C..A.AAC.AG.CTG.TCG
NC_007236	*G. gallus*	......C..........TC......CC.CCCA.C.CAC.TGTATC.
NC_002805	*R. nigromaculata*	T.....C.....GT....AATC.ACTCAC.CCAACCA.CGC.AGGG
NC_001606	*C. carpio*	.......C....G......C...C.TCC.GC.AC..TT.G.TGTC.

The primers for vertebrates, *12SV5F* and *12SV5R,* were modified by the addition of specific tags on the 5′ end to allow the assignment of sequence reads for the relevant sample (Valentini *et al.* 2009a). All of the PCR products were tagged identically on both ends. These tags were composed of CC on the 5′ end followed by seven variable nucleotides that were specific to each sample. The seven variable nucleotides were designed using the *oligoTag* program (http://www.prabi.grenoble.fr/trac/OBITools) to have at least three differences among the tags, to contain no homopolymers longer than two and to avoid a C on the 5′ end so as to allow the detection of a possible deletion within the tag. All of the PCR products from the different samples were first purified using the MinElute PCR purification kit (QIAGEN GmbH), titrated using capillary electrophoresis (QIAxel; QIAgen GmbH) and finally mixed together in equimolar concentration before sequencing.

DNA sequencing

The sequencing was carried out on the Illumina Genome Analyzer IIx (Illumina Inc.), using the Paired-End Cluster Generation Kit V4 and the Sequencing Kit V4 (Illumina Inc.), following the manufacturer's instructions. A total of 108 nucleotides were sequenced on each extremity of the DNA fragments.

Sequence analysis and taxon assignation

The sequence reads were analysed separately with and without the blocking oligonucleotide, using the OBITools (http://www.prabi.grenoble.fr/trac/OBITools). First, the direct and reverse reads corresponding to a single molecule were aligned and merged using the *solexaPairEnd* program, taking into account data quality during the alignment and the consensus computation. Primers and tags were then identified using the *ngsfilter* program. Only sequences with a perfect match on tags and a maximum of two errors on primers were recorded for the subsequent analysis. The amplified regions, excluding primers and tags, were kept for further analysis. Strictly, identical sequences were clustered together using the *obiuniq* program, keeping the information about their distribution among samples. Sequences shorter than 60 bp, or containing ambiguous nucleotides, or with occurrence lower or equal to 100 were excluded using the *obigrep* program. Taxon assignation was achieved using the *ecoTag* program (Pegard *et al.* 2009). *EcoTag* relies on a dynamic programming global alignment algorithm (Needleman & Wunsch 1970) to find highly similar sequences in the reference database. This database was built by extracting the relevant part of the mitochondrial 12S gene from EMBL nucleotide library using the *ecoPCR* program (Bellemain *et al.* 2010; Ficetola *et al.* 2010). A unique taxon was assigned to each unique sequence. This unique taxon corresponds to the last common ancestor node in the NCBI taxonomic tree of all the taxids of the sequences of the reference database that matched against the query sequence. Automatically assigned taxonomic identification was then manually curated to further eliminate those sequences that were the likely result of PCR artefacts (including chimeras, primer dimers or nuclear pseudogenes) or from obvious contaminations. Usually, chimeras can be easily identified by their low identity (<0.9) over the entire query sequence length with any known sequence and by their low frequency when compared with the main prey items. Finally, the prey items were tentatively identified by correlating sequence data with the potential leopard cat vertebrate prey known to be present in the two regions where the faeces were collected, with the constraint that such potential prey must be phylogenetically close to the prey identified in the public database by the *ecoTag* program. The significance of diet differences between ANP and CGNP was assessed by Pearson's chi-squared tests with simulated *P*-values based on 10^6 replicates, using the frequency of occurrence of prey in faeces. Results of such a test have to be analysed carefully because categories used in the contingency table are prey and several prey are detected in each faeces (Wright 2010). This potentially induced a bias if we consider that two prey in the same faeces cannot be considered as independently sampled. If it really exists, the dependency between prey count leads us to overestimate the true number of degrees of freedom. This is a main problem if the test is not rejecting the null hypothesis, but in case of the rejection of this null hypothesis, this places us on the conservative side of the decision.

Rarefaction analysis of prey in faeces originating from ANP and CGNP

We used species rarefaction curve to estimate the total number of prey species likely to be eaten by the leopard cat in the two study areas. The species accumulation, based on the faecal samples, was computed using the analytical formulas of Colwell *et al.* (2004) in EstimateS (Version 8.2, R. K. Colwell, http://purl.oclc.org/estimates).

Results

Of 181 putative felid faeces collected in the field, 38 samples were confirmed to be that of leopard cat with species-specific primers (22 from ANP of 111, and 16 from CGNP of 70). The next-generation sequencing generated about 0.6 and 0.5 million sequences for the samples without and with the blocking oligonucleotide (Table 3), respectively. After applying different filtering programmes, we finally obtained 232 and 141 sequences from the run without and with blocking oligonucleotides, respectively. Sequences within a sample having either a low frequency (e.g. <0.01 when compared with the most frequent sequence) or being very similar to a highly represented sequence were considered to be amplification/sequencing errors and were discarded. All faeces identified as leopard cat with the species-specific primers were confirmed by sequencing. The leopard cat sequence (accession numbers FR873685 and FR873686) was found with a frequency superior to 0.5 in all samples when using only the *12SV5* primer pair (Fig. 3). As in similar experiments (e.g. Deagle *et al.* 2009), we found some human contaminations corresponding to 0.2% and 5.4% of the sequences without and with the blocking oligonucleotide, respectively. A few PCR artefacts with very short sequences were also observed when using the blocking oligonucleotide, but not without blocking.

Effect of blocking oligonucleotide on predator/prey amplification

When amplifications were carried out only with *12SV5* primers, sequences of the leopard cat represented 91.6% of the total count, eight samples (sample 1–8; Fig. 3) exclusively yielded the leopard cat sequence, and 11 different prey taxa were observed in the diet. The blocking oligonucleotide *PrioB* drastically reduced the amplification of the leopard cat sequences, down to 2.2% of the total sequence count, with no leopard cat sequences observed in 31 samples. Under blocking nucleotide conditions, we recorded the amplification of seven additional prey items not previously detected when the same samples were amplified using the *12SV5* primers. The amplification failed in three samples when using the blocking oligonucleotide. The comparison of amplifications without and with blocking oligonucleotide is shown in Fig. 3.

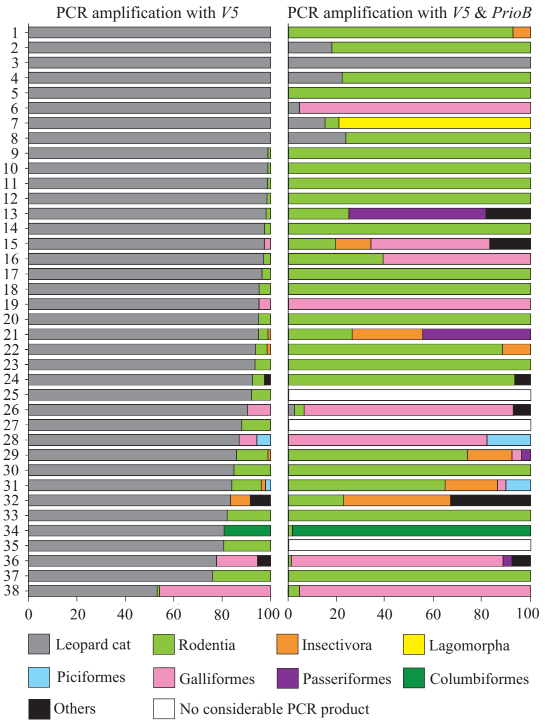


Fig. 3 Comparison of the amplifications of leopard cat and its prey sequences with *12SV5* primers for vertebrates without and with blocking oligonucleotide. The prey items are shown up to the order rank; fish and amphibians are grouped together in the ''others'' category. Each horizontal bar corresponds to the analysis of a single faeces using the *12SV5* primers, either without blocking oligonucleotide (on the left) or with blocking oligonucleotide (on the right). On each bar, the different colours represent the sequence count (%) of predator and prey items present in the sample. Samples 25, 27 and 35 did not show any considerable PCR products with blocking oligonucleotide amplification.

Table 3 Overview of the sequence counts at different stages of the analysis

Primer pair used	*12SV5F/12SV5R*	*12SV5F/12SV5R/PrioB*
Number of properly assembled sequences*	592 648	498 595
Number of unique sequences	44 441	73 414
Number of unique sequences, longer than 60 bp	44 066	46 765
Number of unique sequences, longer than 60 bp, with occurrence in the whole data set higher or equal to 100 (corresponding percentage of properly assembled sequences*)	232 (56.91%)	141 (44.84%)

*Direct and reverse sequence reads corresponding to a single DNA molecule were aligned and merged, producing what we called a ''properly assembled sequence''.

Diet composition of leopard cat

A total of 18 different prey taxa were identified in the diet of the leopard cat, seven of which were identified without ambiguity up to species level (Table 4). A maximum of seven prey items were observed within the same faeces sample, while 15 samples had only a single prey. We were not able to recover any prey DNA from only a single faeces: the experiments without and with blocking oligonucleotide with that sample produced only leopard cat sequences.

The diet composition of the leopard cat from ANP was eclectic; we observed 15 different prey taxa in 22 faeces samples. The house rat predominated the diet (in 68% of the faeces), followed by Asiatic white-toothed shrew (32%) and murree hill frog (27%). We observed seven prey items (Himalayan wood mouse, Kashmir flying squirrel, murree vole, Asiatic white-toothed shrew, chicken, kalij pheasant and jungle crow) within a single faeces, whereas six faeces indicated only a single prey. Overall, Rodentia dominates the diet at ANP with a presence in 91% of the faeces (Fig. 4a). Table 5 gives an overview of the leopard cat diet in Pakistan compared with previous studies.

Eight prey taxa were identified in 16 faeces from CGNP. The house rat predominated the diet (in 44% of the faeces), followed by Kashmir flying squirrel (31%) and Himalayan wood mouse (19%). Rodentia with five different prey species also dominated the diet at CGNP with a presence in 81% of the faeces (Fig. 4b).

While the leopard cat diet in both ANP and CGNP is composed mainly of rodents, the differences between these two areas were significant, both when considering all prey species independently (*P*-value: 0.01; χ^2 test with simulated *P*-value based on 10^6 replicates) and when grouping prey according to their taxonomy (Rodentia, Insectivora, Lagomorpha, Aves, Batracia and Teleostei; *P*-value: 0.03; χ^2 test with simulated *P*-value based on 10^6 replicates). As discussed in the study by Wright (2010), using Pearson chi-squared test for such data can lead to misinterpretation because of the overestimation of the degrees of freedom. By overestimating the degrees of freedom, it is more difficult to reject the null hypothesis. Consequently, rejecting the null hypothesis, as we did, places us on the conservative side of the decision.

Results of the rarefaction analysis are presented in Fig. 5. The number of prey species expected in the pooled faecal samples, based on the rarefaction curve, was 15 (95% CI: 13.91–16.09) and 8 (95% CI: 4.14–11.86) for the ANP and CGNP, respectively. In the case of ANP, 13 of 15 species with a cumulative frequency of 93% in the diet were detected in the first 11 samples. In CGNP, all of the documented prey species were detected in first 13 samples and the rest of the samples reflected their repeats.

Discussion

The leopard cat diet

All documented studies, including the present study, suggest that the order Rodentia is the primary prey base for the leopard cat (presence in 81.2–96.0% of the faeces in six studies, Table 5). Within *Rodentia*, the Muridae family dominates, with a presence in 50.0–86.4% of the faeces in Pakistan and up to 96% in other localities. The arboreal behaviour of the leopard cat (Nowell & Jackson 1996) broadens its trophic niche by enabling it to hunt tree-nesting birds and even flying squirrels in Pakistan. Birds and herpetofauna (reptiles and amphibians) are apparently the other main food groups after mammals. Birds have been reported in all studies, although the highest frequency was observed in Pakistan (presence in 18.7–45.5% of the faeces). In contrast to previous studies, where conventional methods did not allow species identification for birds, we are reporting eight distinct taxa. This specificity is an evident advantage of DNA-based diet methods recently developed. We also report fish in the diet, which have only once been reported previously (Inoue 1972). Our method did not allow the detection of invertebrates or plants, although these have been reported in other studies.

The results of the rarefaction analysis show the efficiency of the molecular method for detecting prey; this is advantageous for studying rare species that inhabit difficult terrains and that do not allow for collecting a large number of samples. Our sample size is smaller than what is generally recommended for classical diet studies; previously, 80 samples have been suggested for common leopards (Mukherjee *et al.* 1994). However, considering the greater detection efficiency of the new method, supported by the rarefaction estimates, our sample size seems to be adequate for estimating the vertebrate diet diversity of the leopard cat in the two studied regions.

The higher diversity of prey detected in samples from ANP as compared to those from CGNP probably reflects the higher productivity and diversity of temperate forests in the former park. The Kashmir flying squirrel prefers to nest on dead trees and is found in both national parks. Its frequency as a prey item was significantly higher in CGNP, the open forests of which probably make flying squirrel more susceptible to predation.

Surprisingly, the leopard cat seems to predate on prey with larger adult body size in Pakistan than in southern parts of its range (Table 5). Larger prey was

Table 4 List of prey taxa found in leopard cat diet (ANP: Ayubia National Park; CGNP: Chitral Gol National Park)

			Number of occurrence		Most similar sequence(s) in public databases using BLAST				Putative taxon identification taking into account the locations where the faeces samples were collected	
MOTU number	Accession number	Number of sequence reads	ANP 22 faeces	CGNP 16 faeces	Species name(s)	Accession number(s)	Query coverage (%)	Maximum identity (%)	Scientific name	Common name
1	FR873673	66680	15	7	*R. tanezumi/rattus*	EU273712/EU273707	100	100	*R. rattus*	House rat
2	FR873674	23746	4	0	*Microtus lusitanicus/pyrenaicus/duodecimcostatus/savii*	AJ972919/AJ972916/AJ972915/AJ972914	100	95	*Hyperacrius wynnei* (?)	Murree vole (?)
3	FR873675	10848	4	0	*Phasianus colchicus/versicolor*	FJ752430/AB164626	100	99	*Lophura leucomelanos* (?)	Kalij pheasant (?)
4	FR873676	10077	2	5	*Eoglaucomys fimbriatus*	AY227562	100	100	*E. fimbriatus*	Kashmir flying squirrel
5	FR873677	9902	5	3	*Apodemus uralensis*	AJ311128	100	100	*Apodemus rusiges*	Himalayan wood mouse
6	FR873678	9827	2	0	*Pucrasia macrolopha*	FJ752429	100	100	*P. macrolopha*	Koklass pheasant
7	FR873679	9361	7	0	*Crocidura gueldenstaedti*	AF434825	97	100	*Crocidura pullata* (?)	Asiatic white-toothed shrew (?)
8	FR873680	8700	0	1	*Columba livia*	GQ240309	100	99	*C. livia* (?)	Rock pigeon (?)
9	FR873681	8469	1	2	*Alectoris chukar*	FJ752426	100	100	*A. chukar*	Chukar partridge
10	FR873682	3626	6	0	*Nanorana parkeri*	AY322333	100	97	*Paa vicina* (?)	Murree hill frog (?)
11	FR873683	3329	0	1	*Lepus* spp.	AY292707	100	94	*Lepus capensis* (?)	Cape hare (?)
12	FR873684	2762	2	0	*Gallus gallus*	GU261719	100	100	*G. gallus*	Chicken
13	FR873687	2049	2	0	Timaliidae	AF376932	100	100	Timaliidae	Babblers
14	FR873688	1770	2	0	*Pica pica; Corvus macrorhynchos/corone/frugilegus/albus*	HQ915867; AB042345/AF386463/Y18522/U38352	100	100	*C. macrorhynchos*	Jungle crow
15	FR873689	1034	2	0	*Picus viridis*	EF027325	100	97	*Dendrocopos* sp. (?)	Woodpecker (?)
16	FR873690	542	0	1	*Dryomys nitedula*	D89005	100	94	*D. nitedula* (?)	Forest dormouse (?)
17	FR873691	434	1	0	*Cephalosilurus apurensis; Liobagrus obesus*	EU179838; DQ321752	100	93	Siluriformes (?)	Cat fish (?)
18	FR873692	105	3	1	*Mus musculus castaneus*	EF108342	100	100	*Mus musculus*	House mouse

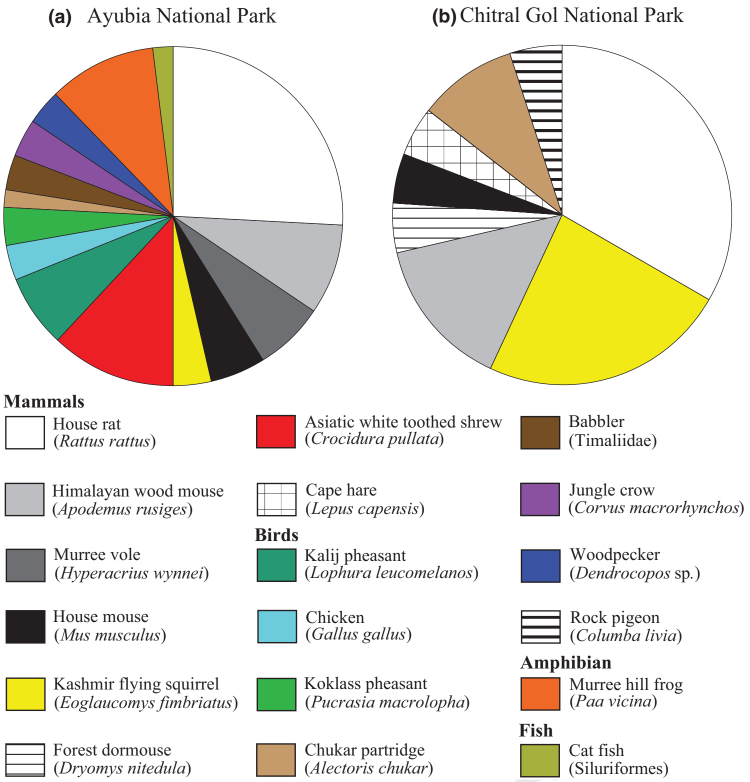


Fig. 4 Composition and comparison of the various prey items consumed and their relative frequency in the diet of the leopard cat at (a) Ayubia National Park and (b) Chitral Gol National Park.

Table 5 Comparison of leopard cat diet across its range in Asia. Except the present study, all other references estimated the diet using traditional morphology-based methods

	Occurrence in faeces, %							
Locality	Rodentia	Other mammals	Birds	Reptiles and amphibians	Fish	Invertebrates	Plant matter	References
ANP, Pakistan	90.9	31.8	45.5	27.3	4.5	Not recorded	Not recorded	Present study
CGNP, Pakistan	81.2	6.2	18.7	0.0	0.0	Not recorded	Not recorded	Present study
Negros-Panay Faunal Region, Philippines	96.0	8.0	8.0	—	—	—	12.0	Fernandez & de Guia (2011)
Khao Yai National Park, Thailand	93.8	24.5	8.2	8.2	—	36.7	—	Austin *et al.* (2007)
Sabah, Malaysian Borneo	93.1	4.2	5.6	19.4	—	11.1	11.1	Rajaratnam *et al.* (2007)
North-central Thailand	89.0	17.0	4.0	—	—	21.0	—	Grassman *et al.* (2005)
Tsushima islands, Japan	91.3	0.3	36.5	22.3	—	24.3	78.8	Tatara & Doi (1994)

usually the house rat (140–280 g), but even bigger prey were occasionally reported. Grassman *et al.* (2005) found remains of Java mouse deer (*Tragulus javanicus*; 1.18–1.28 kg from Weathers & Snyder (1977) and Endo *et al.* 2002) in leopard cat faeces, and Austin *et al.* (2007) once recorded a large ungulate (*Cervus unicolor*; 70.5–112 kg from Idris *et al.* 2000). In Pakistan, many large prey were found in the diet, including the Kashmir flying squirrel (560–734 g; Hayssen 2008), the cape hare (2.10–2.30 kg; Lu 2000), the chukar partridge (450–800 g; del Hoyo *et al.* 1994), the kalij pheasant (564–1150 g; del Hoyo *et al.* 1994), the koklass pheasant

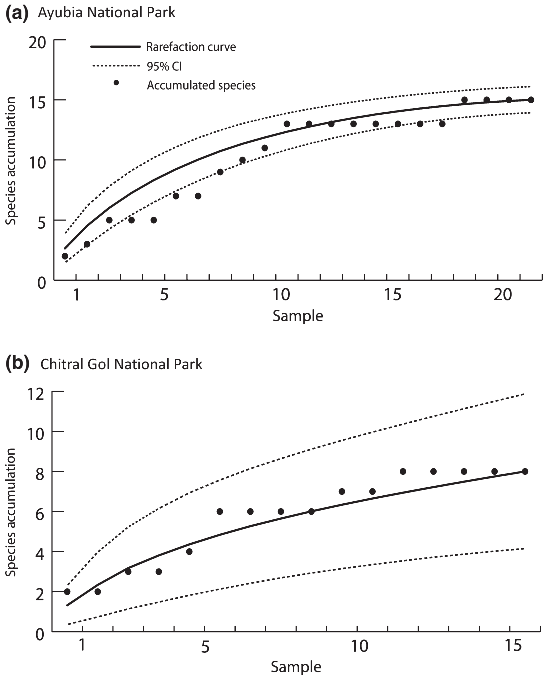


Fig. 5 Species accumulation curves based on the prey species identified in the faecal samples of leopard cat collected in ANP (a) and CGNP (b).

(930–1415 g; del Hoyo *et al.* 1994) and the jungle crow (570–580 g; Matsubara 2003). Two nonexclusive explanations can be proposed to explain such a diet shift towards larger species. First, only juveniles of the larger species may have been captured. It is important to note that remains of juveniles might be difficult to identify in faeces using traditional approaches. DNA-based methods allow straightforward taxon identification, but obviously not the age of prey. Second, the body size of the leopard cat in Pakistan might be larger than in southern areas of its distribution range, possibly explaining their ability to catch larger prey. This last hypothesis tends to be supported by the fact that the leopard cat is known to show considerable variation in size across its geographic distribution, with larger animals in China and Russia (Sunquist & Sunquist 2009), but cannot be confirmed because of the scarcity of data in Pakistan.

We conclude that the results of the present study are in general agreement with previous diet studies of the leopard cat indicating a very eclectic diet. However, the present study highlighted a possible broadening of the diet to include larger prey and provided more precise information by resolving major diet groups to a lower taxonomical level, which was not previously possible using conventional methods.

Conservation implications

The current extent of occurrence of the leopard cat in Pakistan is not resolved (Sheikh & Molur 2004). It historic range started from Chitral and extended to the eastern border of Pakistan, including areas of Swat, Hazara and Ayubia National Park (Nowell & Jackson 1996; Roberts 2005). In the north, it occupied parts of Gilgit Baltistan probably up to an elevation of 3000 m (Habibi 1977). The present study documents its current occurrence in two extremities of its historic range. A leopard cat was photographed in CGNP (SLT 2008), and authors have collected evidence of its presence in Machiara National Park, Azad Jammu and Kashmir, and western parts of the Gilgit Baltistan. This evidence suggests that the historic range of the cat in Pakistan is probably intact, although its population status needs to be determined.

Among the 18 taxa eaten by the cat in Pakistan, four (*Apodemus rusiges*, *Dryomys nitedula*, *Eoglaucomys fimbriatus* and *Lepus capensis*) are categorized as vulnerable (Sheikh & Molur 2004). Because the leopard cat is highly adaptable and appears to be widespread in Pakistan, it may be a potential threat to these species, which have a cumulative frequency of 44.7% of occurrence in faeces. A population assessment of the leopard cats is needed to evaluate the magnitude of this possible threat and to tailor an appropriate management strategy for both prey and predator.

A DNA-based approach for studying carnivore diet

Diet analysis combining next-generation sequencing and vertebrate primers with blocking oligonucleotides has tremendous potential for large-scale studies on carnivore diet. This approach is very robust and presents the complete diet profile of the vertebrate prey consumed. It is highly accurate and discriminates between closely related species in most of the cases. Moreover, a priori knowledge of prey items consumed is not essential, as it is when designing more specific DNA-based approaches. However, such analyses can yield a substantial amount of artefactual sequences including chimeras, nuclear pseudogenes and primer dimers, especially when using the blocking oligonucleotide. As our primers target highly conserved DNA regions in vertebrates, it seems unlikely that a nuclear pseudogene will better match with the *12SV5* primers than the true mitochondrial copies. Furthermore, as mitochondrial copies are much more frequent than nuclear copies, the number of occurrences of any pseudogene sequence should be much lower than the corresponding mitochondrial sequence. With regard to these possible artefacts, we recommend keeping stringent PCR conditions as described in the Materials and Methods section and treating as significant only sequences showing a strong correspondence with a known sequence (at least >0.9) together with a relatively high frequency.

An ongoing debate on DNA-based diet studies concerns the quantification of different prey items consumed and their relative presence in sequence counts. This issue has been highlighted in several recent DNA-based dietary studies (e.g. Deagle *et al.* 2009, 2010; Soininen *et al.* 2009; Valentini *et al.* 2009a). The sequence count cannot be interpreted as quantitative for a few reasons. Biased amplification of some species has been observed when PCR was carried out of a known mixture (Polz & Cavanaugh 1998). Strong biases will occur in dietary studies when primers mismatch with certain prey sequences, resulting in the amplification inclined towards the perfect matches. The two highly conserved regions targeted by the primers *12SV5F* and *12S V5R* make the approach less susceptible to PCR biases. Deagle *et al.* (2010) suggested that differences in the density of mitochondrial DNA in tissues can also bias the sequence count. In the present study, we avoided quantitative interpretations from the results of our sequence counts and recorded only the presence/absence of the different prey in the different faeces.

The blocking oligonucleotide approach has considerable potential for its use in trophic analyses. The design of a blocking oligonucleotide specific to the leopard cat requires knowing the leopard cat sequence for the target DNA region. In this study, the blocking oligonucleotide technique not only inhibited the amplification of the leopard cat DNA, but also uncovered seven more prey taxa in the diet that had not been amplified previously without the blocking oligonucleotide. We used a high concentration of *PrioB* (2 μM) compared with *12SV5F* and *12SV5R* primers (0.1 μM each). For each faeces sample, we systematically ran amplifications without and with blocking oligonucleotide, as amplification with such a relatively high *PrioB* concentration might fail.

One limitation of the approach with the *12SV5F* and *12SV5R* primers proposed here is that it only identifies vertebrate prey. Many carnivores have a more diverse diet, including invertebrates and plants. For example, the Eurasian badger (*Meles meles*) exploits a wide range of food items, especially earthworms, insects and grubs. It also eats small mammals, amphibians, reptiles and birds as well as roots and fruits (Revilla & Palomares 2002). For instance, to study the badger's diet, we suggest complementing the primers for vertebrates with several additional systems, such as primers targeting plant taxa (e.g. Taberlet *et al.* 2007; Valentini *et al.* 2009a) or earthworms (Bienert *et al.* 2012).

One more limitation of this approach for identifying vertebrates is that cases of cannibalism cannot be detected. In such a situation, the predator DNA cannot be distinguished from the prey DNA that belongs to the same species. This limitation was not acknowledged in previous DNA-based diet analyses for vertebrate predators, despite the cases of cannibalism have been documented, for example, in Otariidae (e.g. Wilkinson *et al.* 2000). However, if cannibalism is important from a behavioural point of view, it represents a marginal phenomenon when studying the diet.

Another potential difficulty concerns species identification. In some cases, we had to combine the best match using public databases together with expert knowledge about the available prey in the location where the faeces were collected. For example, in our study, the best match (99%) for MOTU number 3 in public databases corresponded to two species of the genus *Phasianus* (*P. colchicus* and *P. versicolor*). These two species are not recorded in ANP, and thus, we identify this MOTU as the closest relative (Huang *et al.* 2009; Shen *et al.* 2010) occurring in ANP, the kalij pheasant (*Lophura leucomelanos*). If the identification of the kalij pheasant seems reliable, some other putative identifications are more problematic, particularly those having a relatively low identities with known sequences in public databases (i.e. *Hyperacrius wynnei*, *Paa vicina*, *L. capensis*, *Dendrocopos* sp., and *D. nitedula*). To remove such uncertainties, we recommend constructing a local reference database when possible.

The results of the present study correspond to summer diet and may not reflect the complete diet profile of the leopard cat in Pakistan. In future, it would be interesting to collect samples throughout the year, with the attendant possibility of revealing more prey taxa than what we have observed in this study.

Conclusion

Noninvasive sampling is the only way to study the diet of elusive animals like the leopard cat. In Pakistan, we obtained results confirming the eclectic characteristics of this predator, together with an extension of the diet towards larger prey. The DNA-based approach has a better resolution than conventional approach-based identification of prey from hair and bone remains. While DNA-based methods cannot assess prey ages, conventional approaches might reveal the potential ages of the prey when necessary, possibly determining whether juveniles or adults of larger prey were consumed. As a consequence, DNA-based diet analysis can provide a valuable complement to conventional methods.

The DNA-based approach we propose here is particularly robust and simple to implement and allows the possibility of very large-scale analyses. It can be applied to other carnivore species with only a slight adjustment concerning the design of the blocking oligonucleotide.

Acknowledgements

The authors are thankful to A. Valentini, D. Rioux and L. Gielly for their valuable suggestions and comments throughout the course of this study. For English correction in the manuscript, we express our gratitude to M. MacFarlane and M. Reynolds. W. Shehzad and T. Riaz were supported by PhD scholarship from the HEC (Higher Education Commission), Government of Pakistan. The Himalayan Wildlife Foundation and Snow Leopard Trust provided financial support for sample collections in Ayubia and Chitral Gol National Parks. The staff of the Khyber Pakhtunkhwa Wildlife Department and Snow Leopard Trust assisted in sample collection.

References

Austin SC, Tewes ME, Grassman Jr LI, Silvy NJ (2007) Ecology and conservation of the leopard can *Prionailurus bengalensis* and clouded leopard *Neofelis nebulosa* in Khao Yai National Park, Thailand. *Acta Zoologica Sinica*, **53**, 1–14.

Bagchi S, Mishra C (2006) Living with large carnivores: predation on livestock by the snow leopard (*Uncia uncia*). *Journal of Zoology*, **268**, 217–224.

Bellemain E, Carlsen T, Brochmann C *et al.* (2010) ITS as DNA barcode for fungi: an in silico approach reveals potential PCR biases. *BMC Microbiology*, **10**, 189.

Bienert R, de Danieli S, Miquel C *et al.* (2012) Tracking earthworm communities from soil DNA. *Molecular Ecology*, 7 **???**, ???–???, in press.

Blankenship LE, Yayanos AA (2005) Universal primers and PCR of gut contents to study marine invertebrate diets. *Molecular Ecology*, **14**, 891–899.

Bradley BJ, Stiller M, Doran-Sheehy DM *et al.* (2007) Plant DNA sequences from feces: potential means for assessing diets of wild primates. *American Journal of Primatology*, **69**, 699–705.

Brownstein MJ, Carpten JD, Smith JR (1996) Modulation of non-templated nucleotide addition by *Taq* polymerase: primer modification that facilitate genotyping. *BioTechniques*, **20**, 1004–1010.

Colwell RK, Mao CX, Chang J (2004) Interpolating, extrapolating, and comparing incidence-based species accumulation curves. *Ecology*, **85**, 2717–2727.

Deagle BE, Tollit DJ (2007) Quantitative analysis of prey DNA in pinniped faeces: potential to estimate diet composition? *Conservation Genetics*, **8**, 743–747.

Deagle BE, Jarman SN, Pemberton D, Gales NJ (2005) Genetic screening for prey in the gut contents from a giant squid (*Architeuthis* sp.). *Journal of Heredity*, **96**, 417–423.

Deagle BE, Eveson JP, Jarman SN (2006) Quantification of damage in DNA recovered from highly degraded samples – a case study on DNA in faeces. *Frontier in Zoology*, **3**, 11.

Deagle BE, Gales NJ, Evans K *et al.* (2007) Studying seabird diet through genetic analysis of faeces: a case study on macaroni penguins (*Eudyptes chrysolophus*). *PLoS ONE*, **2**, e831.

Deagle BE, Kirkwood R, Jarman SN (2009) Analysis of Australian fur seal diet by pyrosequencing prey DNA in faeces. *Molecular Ecology*, **18**, 2022–2038.

Deagle BE, Chiaradia A, McInnes J, Jarman SN (2010) Pyrosequencing faecal DNA to determine diet of little penguins: is what goes in what comes out? *Conservation Genetics*, **11**, 2039–2048.

Dunshea G (2009) DNA-based diet analysis for any predator. *PLoS ONE*, **4**, e5252.

Endo H, Kimura J, Sasaki M *et al.* (2002) Functional morphology of the mastication muscles in the lesser and greater mouse deer. *Journal of Veterinary Medical Science*, **64**, 901–905.

Farooque M (2007) *Management Plan of Ayubia National Park*. Khyber Pakhtunkhwa Wildlife Department, Peshawar.

Fernandez DAP, de Guia APO (2011) Feeding habits of Visayan leopard cats (*Prionailurus bengalensis rabori*) in sugarcane fields of Negros Occidental, Philippines. *Asia Life Sciences*, **20**, 143–154.

Ficetola GF, Coissac E, Zundel S *et al.* (2010) An *in silico* approach for the evaluation of DNA barcodes. *BMC Genomics*, **11**, 434.

Gaines WL (2001) Large carnivore surveys in the North Karakorum Mountains, Pakistan. *Natural Areas Journal*, **21**, 168–171.

GoNWFP & IUCN (1996) *Sarhad Provincial Conservation Strategy*. Sarhad Programme Office, IUCN–The World Conservation Union, Peshawar.

Grassman LI, Tewes ME, Silvy NJ, Kreetiyutanont K (2005) Spatial organization and diet of the leopard cat (*Prionailurus bengalensis*) in north-central Thailand. *Journal of Zoology*, **266**, 45–54.

Green SJ, Minz D (2005) Suicide polymerase endonuclease restriction, a novel technique for enhancing PCR amplification of minor DNA templates. *Applied and Environmental Microbiology*, **71**, 4721–4727.

Habibi K (1977) *The Mammals of Afghanistan: Their Distribution and Status*. UNDP, FAO and Ministry of Agriculture, Kabul.

Hayssen V (2008) Patterns of body and tail length and body mass in Sciuridae. *Journal of Mammalogy*, **89**, 852–873.

Höss M, Kohn M, Pääbo S, Knauer F, Schröder W (1992) Excrement analysis by PCR. *Nature*, **359**, 199.

del Hoyo J, Elliot A, Sargatal J (1994) *Handbook of the Birds of the World*, Vol. 2. New World Vultures to Guinea Fowl Lynx Edicions, Barcelona.

Huang ZH, Liu NF, Xiao YA *et al.* (2009) Phylogenetic relationships of four endemic genera of the Phasianidae in China based on mitochondrial DNA control-region genes. *Molecular Phylogenetics and Evolution*, **53**, 378–383.

Idris I, Moin S, Sulah S, Jiwan D (2000) Some physical characteristics of Sambar deer (*Cervus unicolor*). *Pertanika Journal of Tropical Agricultural Science*, **33**, 55–59.

Inoue T (1972) The food habit of Tsushima leopard cat, *Felis bengalensis* spp., analysed from their scats. *Journal of the Mammalogical Society of Japan*, **5**, 155–169.

Izawa M, Doi T (1991) Status of conservation and management of two species of felidae in Japan (in Japanese). *Mammalian Science*, **31**, 15–22.

Jarman SN, Gales NJ, Tierney M, Gill PC, Elliott NG (2002) A DNA-based method for identification of krill species and its application to analysing the diet of marine vertebrate predators. *Molecular Ecology*, **11**, 2679–2690.

Jarman SN, Deagle BE, Gales NJ (2004) Group-specific polymerase chain reaction for DNA-based analysis of species diversity and identity in dietary samples. *Molecular Ecology*, **13**, 1313–1322.

Jarman SN, Redd KS, Gales NJ (2006) Group-specific primers for amplifying DNA sequences that identify Amphipoda, Cephalopoda, Echinodermata, Gastropoda, Isopoda, Ostracoda and Thoracica. *Molecular Ecology Notes*, **6**, 268–271.

Kageyama T, Sato Y, Nishizawa S, Teramae N (2008) Competitive binding of small ligands to nucleobases in AP site-containing DNA duplexes. *Nucleic Acids Symposium Series*, **52**, 119–120.

King RA, Vaughan IP, Bell JR, Bohan DA, Symondson WOC (2010) Prey choice by carabid beetles feeding on an earthworm community analysed using species- and lineage-specific PCR primers. *Molecular Ecology*, **19**, 1721–1732.

Kohn MH, Wayne RK (1997) Facts from feces revisited. *Trends in Ecology & Evolution*, **12**, 223–227.

Krahn MM, Herman DP, Matkin CO *et al.* (2007) Use of chemical tracers in assessing the diet and foraging regions of eastern North Pacific killer whales. *Marine Environmental Research*, **63**, 91–114.

Lau MWN, Fellowes JR, Chan BPL (2010) Carnivores (Mammalia: Carnivora) in South China: a status review with notes on the commercial trade. *Mammal Review*, **40**, 247–292.

Li MJ, Sato Y, Nishizawa S *et al.* (2009) 2-aminopurine-modified abasic-site-containing duplex DNA for highly selective detection of theophylline. *Journal of the American Chemical Society*, **131**, 2448–2449.

Liles MR, Manske BF, Bintrim SB, Handelsman J, Goodman RM (2003) A census of rRNA genes and linked genomic sequences within a soil metagenomic library. *Applied and Environmental Microbiology*, **69**, 2684–2691.

Lovari S, Boesi R, Minder I *et al.* (2009) Restoring a keystone predator may endanger a prey species in a human-altered ecosystem: the return of the snow leopard to Sagarmatha National Park. *Animal Conservation*, **12**, 559–570.

Lu X (2000) Body weights of the Cape hare *Lepus capensis* in the northern China. *Acta Theriologica*, **45**, 271–280.

Macdonald DW, Loveridge AJ, Nowell K (2010) Dramatis personae: an introduction to the wild felids. In: *Biology and Conservation of Wild Felids* (eds Macdonald DW and Loveridge AJ). Oxford University Press, New York. 8

Magnuson VL, Ally DS, Nylund SJ *et al.* (1996) Substrate nucleotide-determinated non-templated addition of adenine by Taq DNA polymerase: implications for PCR-based genotyping and cloning. *BioTechniques*, **21**, 700–709.

Matsubara H (2003) Comparative study of territoriality and habitat use in syntopic Jungle Crow (*Corvus macrorhynchos*) and Carrion Crow (*C. corone*). *Ornithological Science*, **2**, 103–111.

Mills LS, Soule ME, Doak DF (1993) The keystone-species concept in ecology and conservation. *BioScience*, **43**, 219–224.

Mirza ZB (2003) *Biological Baseline Study of Chitral Gol National Park*. Protected Areas Management Project, Islamabad.

Mitani N, Mihara S, Ishii N, Koike H (2009) Clues to the cause of the Tsushima leopard cat (*Prionailurus bengalensis euptilura*) decline from isotopic measurements in three species of Carnivora. *Ecological Research*, **24**, 897–908.

Mukherjee S, Goyal SP, Chellam R (1994) Standardization of scat analysis techniques for leopard (*Panthera pardus*) in Gir National Park, western India. *Mammalia*, **58**, 139–143.

Mukherjee S, Krishnan A, Tamma K *et al.* (2010) Ecology driving genetic variation: a comparative phylogeography of jungle cat (*Felis chaus*) and leopard cat (*Prionailurus bengalensis*) in India. *PLoS ONE*, **5**, 16.

Natoli E (1990) Mating strategies in cats – a comparison of the role and importance of infanticide in domestic cats, *Felis catus* L and lions, *Panthera leo* L. *Animal Behaviour*, **40**, 183–186.

Needleman SB, Wunsch CD (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, **48**, 443–453.

Nowell K, Jackson P (1996) *Wild Cats, Status Survey and Conservation Action Plan*. IUCN, ????, Switzerland, 382p. 9

Oli MK, Taylor IR, Rogers ME (1994) Snow leopard (*Panthera Uncia*) predation of livestock – an assessment of local perceptions in the Annapurna Conservation area, Nepal. *Biological Conservation*, **68**, 63–68.

Pegard A, Miquel C, Valentini A *et al.* (2009) Universal DNA based methods for assessing the diet of grazing livestock and wildlife from faeces. *Journal of Agricultural and Food Chemistry*, **57**, 5700–5706.

Polz MF, Cavanaugh CM (1998) Bias in template-to-product ratios in multitemplate PCR. *Applied and Environmental Microbiology*, **64**, 3724–3730.

Power ME, Tilman D, Estes JA *et al.* (1996) Challenges in the quest for keystones. *BioScience*, **46**, 609–620.

Rajaratnam R, Sunquist M, Rajaratnam L, Ambu L (2007) Diet and habitat selection of the leopard cat (*Prionailurus bengalensis borneoensis*) in an agricultural landscape in Sabah, Malaysian Borneo. *Journal of Tropical Ecology*, **23**, 209–217.

Revilla E, Palomares F (2002) Does local feeding specialization exist in Eurasian badgers? *Canadian Journal of Zoology-Revue Canadienne De Zoologie*, **80**, 83–93.

Rho P (2009) Use of GIS to develop a multivariate habitat model for the leopard cat (*Prionailurus bengalensis*) in mountainous region of Korea. *Journal of Ecology & Field Biology*, **32**, 229–236.

Riaz T, Shehzad W, Viari A *et al.* (2011) ecoPrimers: inference of new DNA barcode markers from whole genome sequence analysis. *Nucleic Acids Research*, **???**, ???–???, doi: 10.1093/nar/gkr732. 10

Roberts TJ (2005) *Field Guide to the Large and Medium Sized Mammals of Pakistan*. Oxford University Press, Oxford, UK.

Scott DM, Gemita E, Maddox TM (2004) Small cats in human modified landscapes in Sumatra. *Cat News*, **40**, 23–25.

Sheikh KM, Molur S (2004) Status and red list of Pakistan's mammals. Based on the Conservation Assessment and Management Plan Workshop. IUCN Pakistan.

Shen YY, Liang L, Sun YB *et al.* (2010) A mitogenomic perspective on the ancient, rapid radiation in the Galliformes with an emphasis on the Phasianidae. *BMC Evolutionary Biology*, **10**, 132.

Snow Leopard Trust (2008) Population monitoring of large carnivores in Chitral Gol National Park. Progress update for 2007–2008, submitted to NWFP Wildlife Department. 11

Soininen EM, Valentini A, Coissac E *et al.* (2009) Analysing diet of small herbivores: the efficiency of DNA barcoding coupled with high-throughput pyrosequencing for deciphering the composition of complex plant mixtures. *Frontiers in Zoology*, **6**, 9.

Sunquist ME, Sunquist FC (2002) *Wild Cats of the World*. University of Chicago Press, Chicago, Illinois, 416pp.

Sunquist ME, Sunquist FC (2009) Family Felidae (cats). In: *Handbook of the Mammals of the World. Vol. 1. Carnivores* (eds Wilson DE and Mittermeier RA), pp. 54–168. Lynx Edicions, Barcelona.

Symondson WOC (2002) Molecular identification of prey in predator diets. *Molecular Ecology*, **11**, 627–641.

Taberlet P, Coissac E, Pompanon F *et al.* (2007) Power and limitations of the chloroplast *trn*L(UAA) intron for plant DNA barcoding. *Nucleic Acids Research*, **35**, e14.

Tatara M, Doi T (1994) Comparative analyses on food-habits of Japanese marten, Siberian weasel and leopard cat in The Tsushima islands, Japan. *Ecological Research*, **9**, 99–107.

Valentini A, Miquel C, Nawaz MA *et al.* (2009a) New perspectives in diet analysis based on DNA barcoding and parallel pyrosequencing: the *trn*L approach. *Molecular Ecology Resources*, **9**, 51–60.

Valentini A, Pompanon F, Taberlet P (2009b) DNA barcoding for ecologists. *Trends in Ecology and Evolution*, **24**, 110–117.

Vestheim H, Jarman SN (2008) Blocking primers to enhance PCR amplification of rare sequences in mixed samples – a case study on prey DNA in Antarctic krill stomachs. *Frontiers in Zoology*, **5**, 11.

Vestheim H, Edvardsen B, Kaartvedt S (2005) Assessing feeding of a carnivorous copepod using species-specific PCR. *Marine Biology*, **147**, 381–385.

Wang YF, Ng MTT, Zhou TY *et al.* (2008) C3-Spacer-containing circular oligonucleotides as inhibitors of human topoisomerase I. *Bioorganic & Medicinal Chemistry Letters*, **18**, 3597–3602.

Watanabe S (2009) Factors affecting the distribution of the leopard cat *Prionailurus bengalensis* on East Asian islands. *Mammal Study*, **34**, 201–207.

Weathers WW, Snyder GK (1977) Hemodynamics of the lesser mouse deer, *Tragulus javanicus*. *Journal of Applied Physiology*, **42**, 679–681.

Wilkinson IS, Childerhouse SJ, Duignan PJ, Gulland FMD (2000) Infanticide and cannibalism in the New Zealand sea lion, *Phocarctos hookeri*. *Marine Mammal Science*, **16**, 494–500.

Wright BE (2010) Use of chi-square tests to analyze scat-derived diet composition data. *Marine Mammal Science*, **26**, 395–401.

Zaidi RH, Jaal Z, Hawkes NJ, Hemingway J, Symondson WOC (1999) Can multiple-copy sequences of prey DNA be detected amongst the gut contents of invertebrate predators? *Molecular Ecology*, **8**, 2081–2087.

T.R., E.C. and P.T. are co-inventors of a pending French patent on the primer pair named *12SV5F* and *12SV5R* and on the use of the amplified fragment for identifying vertebrate species from environmental samples. This patent only restricts commercial applications and has no impact on the use of this method by academic researchers.

Data accessibility

DNA sequences of the V5 loop of the mitochondrial 12S gene: GenBank accessions FR873673–FR873692.

Fasta file and filtered data deposited in the Dryad repository: doi: 10.5061/dryad.443t4m1q.

Annex B
Article 4

MOLECULAR ECOLOGY
RESOURCES

New environmental metabarcodes for analysing soil DNA: potential for studying past and present ecosystems

Journal:	*Molecular Ecology Resources*
Manuscript ID:	Draft
Manuscript Type:	Resource Article
Date Submitted by the Author:	n/a
Complete List of Authors:	Epp, Laura; University of Oslo, NCB - National Centre for Biosystematics, Natural History Museum Boessenkool, Sanne; University of Oslo, NCB - National Centre for Biosystematics, Natural History Museum Bellemain, Eva; University of Oslo, NCB - National Centre for Biosystematics, Natural History Museum Haile, James; Murdoch University, Ancient DNA laboratory Esposito, Alfonso; Free University of Bozen, Faculty of Science and Technology Riaz, Tiayyba; Universite de Grenoble, Laboratoire d'Ecologie Alpine Erséus, Christer; University of Gothenburg, Department of Zoology Gusarov, Vladimir; National Centre for Biosystematics, Natural History Museum, University of Oslo Edwards, Mary; University of Southampton, School of Geography Johnsen, Arild; University of Oslo, NCB - National Centre for Biosystematics, Natural History Museum Stenøien, Hans; Museum of Natural History and Archaeology, Norwegian University of Science and Technology, Systematics and Evolution Group, Section of Natural History Hassel, Kristian; Museum of Natural History and Archaeology, Norwegian University of Science and Technology, Systematics and Evolution Group, Section of Natural History Kauserud, Havard; University of Oslo, Department of Biology, Microbial Evolution Research Group (MERG) Yoccoz, Nigel; University of Tromsø, Arctic and Marine Biology Brathen, Kari-Anne; University of Tromsø, Arctic and Marine Biology Willerslev, Eske; University of Copenhagen, Centre for GeoGenetics Taberlet, Pierre; Université Joseph Fourier, Laboratoire d'Ecologie Alpine, CNRS UMR 5553; Coissac, Eric; Universite de Grenoble, Laboratoire d'Ecologie Alpine Brochmann, Christian; University of Oslo, NCB - National Centre for Biosystematics, Natural History Museum
Keywords:	environmental DNA, primers, metabarcoding, ancient DNA, Arctic

New environmental metabarcodes for analysing soil DNA: potential for studying past and present ecosystems

Laura S. Epp[1*]; Sanne Boessenkool[1*]; Eva P. Bellemain[1]; James Haile[2,3]; Alfonso Esposito[1§]; Tiayyba Riaz[4]; Christer Erséus[5]; Vladimir Gusarov[1]; Mary E. Edwards[6]; Arild Johnsen[1]; Hans K. Stenøien[7]; Kristian Hassel[7]; Håvard Kauserud[8]; Nigel G. Yoccoz[9]; Kari Anne Bråthen[9]; Eske Willerslev[2]; Pierre Taberlet[4]; Eric Coissac[4#] & Christian Brochmann[1#]

[1]National Centre for Biosystematics, Natural History Museum, University of Oslo, P.O. Box 1172, Blindern, NO-0318 Oslo, Norway

[2]The Centre of Excellence for GeoGenetics, Natural History Museum of Denmark, Øster Voldgade 5-7, 1350 Copenhagen K, Denmark

[3]Ancient DNA Research Laboratory, Murdoch University, South Street, Perth, 6150 Australia

[4]Laboratoire d'Ecologie Alpine, Université Joseph Fourier, BP 53, 2233 Rue de la Piscine, 38041 Grenoble Cedex 9, France

[5]Department of Zoology, University of Gothenburg, Box 463, SE-405 30 Göteborg, Sweden

[6]Geography and Environment, University of Southampton, University Road, Southampton, SO17 1BJ, United Kingdom

[7]Systematics and Evolution Group, Museum of Natural History and Archeology, Norwegian University of Science and Technology, N-7491 Trondheim, Norway

[8]Microbial Evolution Research Group (MERG), Department of Biology, University of Oslo, PO Box 1066 Blindern, N-0316, Oslo, Norway

[9]*Department of Arctic and Marine Biology, University of Tromsø, NO-9037 Tromsø, Norway*

*contributed equally

#shared senior authorship

§present address: *Faculty of Science and Technology, Free University of Bozen, Sernesistrasse,1, I-39100 Bozen, Italy*

Correspondence: Laura Epp, National Centre for Biosystematics, Natural History Museum, University of Oslo, P.O. Box 1172 Blindern, NO-0318 Oslo, Norway. Fax. +47 22851835. Email: laura.epp@nhm.uio.no

Keywords: environmental DNA; primers; metabarcoding; ancient DNA; Arctic;

Running title: Metabarcodes to analyse soil DNA

Abstract

Metabarcoding approaches use total and typically degraded DNA from environmental samples to analyse biotic assemblages and can potentially be carried out for any kinds of organisms in an ecosystem. These analyses rely on specific markers, here called metabarcodes, which should be optimized for taxonomic resolution, minimal bias in amplification of the target organism group and short sequence length. Using bioinformatic tools, we developed metabarcodes for several groups of organisms: fungi, bryophytes, enchytraeids, beetles and birds. The ability of these metabarcodes to amplify the target groups was systematically evaluated by (1) *in silico* PCRs using all standard sequences in the EMBL public database as templates, (2) *in vitro* PCRs of DNA extracts from surface soil samples from a site in Varanger, northern Norway, and (3) *in vitro* PCRs of DNA extracts from permanently frozen sediment samples of late-Pleistocene age (~ 16 000–50 000 yr BP) from two Siberian sites, Duvanny Yar and Main River. Comparison of the results from the *in silico* PCR with those obtained *in vitro* showed that the *in silico* approach offered a reliable estimate of the suitability of a marker. All target groups were detected in the environmental DNA, but we found large variation in the level of detection among the groups and between modern and ancient samples. Success rates for the Pleistocene samples were highest for fungal DNA, whereas bryophyte, beetle and bird sequences could also be retrieved, but to a much lesser degree. The metabarcoding approach has considerable potential for biodiversity screening of modern samples and also as a paleoecological tool.

Introduction

Sequencing of environmental DNA retrieved from soils and sediments plays an important role in the efforts to explore the biodiversity of prokaryotes (Stackebrandt *et al.* 1993; Dunbar *et al.* 1999; Rappé & Giovannoni 2003). The targeted retrieval of DNA from environmental samples also promises great potential for the study of eukaryote biodiversity of recent as well as past environments (Hofreiter *et al.* 2003; Willerslev *et al.* 2003; Willerslev *et al.* 2007; Coolen *et al.* 2009; Sønstebø *et al.* 2010). In particular, DNA from vascular plants (Sønstebø *et al.* 2010), mammals (Haile *et al.* 2009) and fungi (Lydolph *et al.* 2005) targeted within total soil DNA has yielded promising results. This approach is particularly interesting for organisms that do not fossilize readily or for which only few fossils are found. However its potential for ecosystem-wide biodiversity and paleoecological reconstructions that include, for example, invertebrates and vertebrates other than mammals, is currently unclear and requires further evaluation.

Extracts from soil contain DNA from organisms living in the soil at the time of sampling as well as DNA from dead cells and DNA deposited from the surrounding environment (Levy-Booth *et al.* 2007). While the fraction derived from live organisms is largely intracellular and intact, the other fractions will be extracellular and probably highly degraded (Pietramellara *et al.* 2009; Valentini *et al.* 2009b). Such degradation accrues over time and, particularly for ancient environmental DNA, analyses are typically restricted to very short fragments from multi-copy loci (Pääbo *et al.* 2004).

Genetic markers suitable for diversity analyses through taxonomic identification of DNA preserved in environmental samples (a form of DNA barcoding *sensu lato*; Valentini *et al.* 2009b) must fulfil requirements which partly differ from those of DNA barcodes used for the identification of single specimens (barcoding *sensu stricto*; Valentini *et al.* 2009b). First, they should be short enough to allow amplification from degraded DNA in environmental samples. Second, a diagnostic DNA sequence that is more or less identical within but variable between species is required for optimal taxonomic resolution. Third, this variable DNA marker has to be flanked by highly conserved stretches to which amplification primers can bind. These priming sites should be conserved enough to amplify DNA from a mixture of species belonging to the target organism group with minimal bias (Bellemain *et al.* 2010). Finally, the amplification primers should be highly specific to the target organism group in order to avoid amplification of non-target DNA preserved in the sample.

Criterion two, taxonomic resolution to the species level, is of paramount importance for barcoding single specimens (Hebert *et al.* 2003), and the standard marker used for animals is a fragment of the mitochondrial cytochrome c oxidase subunit I gene (COI) with a length of 648 bp. For the analysis of degraded DNA from environmental samples, criteria one, three and four are the most important ones (Valentini *et al.* 2009b) and the long COI fragment is therefore not optimal. Primers to target a shorter fragment of the COI gene in all major eukaryotic groups have been suggested (Meusnier *et al.* 2008), but these primers are not conserved even within vertebrates (Ficetola *et al.* 2010), and the originally proposed primer set was not used in consecutive studies (Hajibabaei *et al.* 2011; Shokralla *et al.* 2011). There is thus a clear need for further development and evaluation of primers and markers for analysis of degraded DNA from environmental samples. With reference to the terms metagenomics and barcoding, we designate barcoding markers specifically designed for environmental samples as “metabarcodes” (Pompanon *et al.* 2011).

In the present study we designed metabarcoding markers and evaluated their potential for studying the biodiversity of different organism groups in past and present arctic ecosystems using DNA from soils and sediments. We targeted a range of phylogenetically and ecologically distinct groups that have received relatively little or no attention in previous studies of (ancient) environmental DNA: bryophytes, enchytraeids, beetles and birds. We also designed a new metabarcoding primer for fungi, which amplifies a somewhat shorter fragment compared with the widely used fungi-specific ITS markers (see Bellemain *et al.* 2010). We selected these taxonomic groups because they are ecologically important and occur frequently in the Arctic (Callaghan *et al.* 2004). Some of them are closely associated with the soil, such as fungi and enchytraeids, while others live above ground, such as birds.

For the animal groups we used mitochondrial DNA, which is well suited for work with degraded samples due to its high copy number per cell (Pääbo *et al.* 2004). Rather than using the standard COI region as a starting point, for the reasons outlined above, we screened either complete mitochondrial genomes or focused on the mitochondrial rRNA genes 12S and 16S. The latter display stem-loop structures leading to a variation of short stretches of highly conserved and stretches of highly variable DNA (Hickson *et al.* 1996; De Rijk *et al.* 1999). Nuclear rRNA genes also contain such structures, which have been shown to be valuable as markers for species identification (Sonnenberg *et al.* 2007; Raupach *et al.* 2010) due to their hyper-variable regions. Short fragments of mitochondrial rRNA genes are good candidate regions for metabarcodes and have previously been identified for vertebrate identification in degraded samples (Riaz *et al.* 2011).

In this study, we used bioinformatic approaches (Ficetola *et al.* 2010; Riaz *et al.* 2011) to first design a set of metabarcodes suitable for detection of these ecologically important groups and evaluate their performance *in silico*. Second, we evaluated the success of our newly designed markers in retrieving DNA of the target organism groups in recent soils (from the Varanger Peninsula in northern Norway) and in frozen late-Pleistocene sediment samples (Main River and Duvanny Yar, northeast Siberia). Arctic permafrost sediments have previously been a main focus of ancient environmental DNA studies (Willerslev *et al.* 2003), as DNA degradation is retarded under cold conditions (Pääbo *et al.* 2004). The Siberian samples range in age from ~16 000 – 50 000 yr BP, a period encompassing the Last Glacial Maximum cold climatic interval (LGM) and characterized by diverse ecosystems with no contemporary (Blinnikov *et al.* 2011) analogues. We compared the predictions obtained from the bioinformatic analyses (*in silico* analyses) with our tests on soil DNA (*in vitro* analyses), and discuss the potential of metabarcoding approaches for the analysis of present and past biodiversity of the different groups studied here.

Material and Methods

Study sites and samples

Recent soil samples were obtained from four pairs of heath and meadow plots used in other ecological studies (Ravolainen *et al.* 2011) and located on the Varanger Peninsula in northern Norway (70º19´ N, 30º01´ E and 70º18´ N, 29º06´ E; 110–290 m a. s. l.). The area is characterized by a mosaic of dwarf shrub heath, herb- and grass-rich meadows and willow thickets. A total of eight samples were analysed from the meadow plots (sample names beginning with ENG, Table 2), and seven samples from the heath plots (sample names beginning with HEI, Table 2). Samples were taken in March 2007 by hammering 15 cm long and 5 cm wide metal cylinders, which had been thoroughly cleaned and treated with sodium hypochlorite to remove DNA prior to use, into frozen soil cleared of surface vegetation and kept frozen until processed for DNA analyses.

Samples from ancient permafrost were obtained from two key paleoecological sites in Eastern Siberia: 1) an exposure at Duvanny Yar, on the Kolyma River, northern Sakha (Yakutia) Republic, Russia (68°40' N, 159°05' E) and 2) an exposure on the Main River, a tributary of the Anadyr River in southern Chukotka, Russia (64°17' N, 171°15' E; Kuzmina *et al*. 2011). At both sites, samples were taken at different depths along the exposures by drilling cores horizontally with equipment that had been cleaned thoroughly and treated with sodium hypochlorite prior to use. The sampled cores were stored intact and frozen until processing. A total of 14 samples were analysed from each site. Organic material (usually plant macrofossils) extracted from the soil/sediment samples by sieving was radiocarbon dated at facilities in Poznan, Poland or Oxford, UK. Two of the Varanger soil samples used here were also radiocarbon dated and both were confirmed to be 'modern' (i.e. from the past several decades, see Table 2).

Design and optimization of metabarcodes

Metabarcoding markers were designed and evaluated using a bioinformatic approach employing the OBITools (www.grenoble.prabi.fr/trac/OBITools). Detailed information about settings and databases used are compiled in the Supplementary Material. Searches of potentially suitable metabarcodes were carried using the program ecoPrimers (Riaz *et al.* 2011). This program uses a defined input database of homogeneous sequences (e.g. full mitochondrial genomes) to search for conserved stretches of DNA suitable to be used as primers that flank a region of a specified length (in this study 20– 500 bp excluding primers, see Supplementary Material). To simulate more realistic PCR conditions (similar to suggestions by Dieffenbach *et al.* 1993), we allowed a maximum of three mismatches between the primer and the target sequences, but no mismatches in the two last bases on the 3' end of the primer. If not otherwise specified (see Supplementary Material), primers were required to strictly (i.e. without errors) match 70% of target sequences (option –q 0.70), to match 90% of target sequences allowing a specified number of mismatches (option –s 0.90), and not to match more than 10% of non-target sequences (option –x 0.10).

For each primer pair, ecoPrimers calculates two quality indices; taxonomic coverage (coverage index B_c) and taxonomic resolution capacity (specificity index B_s), as defined by Ficetola *et al.* (2010). Taxonomic coverage is the number of amplified target species relative to the total number of target species in the input database. Taxonomic resolution capacity is the number of unambiguously identified species relative to the total number of amplified target species. From the ecoPrimers output we selected primer pairs with the highest taxonomic coverage and resolution capacity and a relatively short amplicon length. Primer characteristics were optimised using the programs Primer3 (http://frodo.wi.mit.edu/primer3/) or FastPCR (Kalendar *et al.* 2009) and by visual inspection of an alignment containing a subset of sequences from the target taxa. The primers were also evaluated using the program ecoPCR (Ficetola *et al.* 2010), which performs *in silico* PCRs on a specified database, such as one compiled from all standard sequences in the EMBL Nucleotide Sequence Database (Cochrane *et al.* 2009). The output of ecoPCR contains a list of all sequence entries matching the respective primer pair in a way allowing PCR amplification. This can be used to calculate the taxonomic resolution capacity and coverage for the target group, as defined above, and primer specificity to the target group (number of target species relative to all amplified species). For all final metabarcoding primers, specificity and taxonomic resolution capacity were evaluated using ecoPCR on a database constructed from the standard sequences in the release 107 of the EMBL database (March 2011) with the following parameters: 1) amplicon lengths between 20 and 1000 base pairs, and 2) a maximum of three mismatches between the primer and the target sequence, but no mismatches in the last two bases on the 3' end. The coverage of finalized primer pairs was calculated by performing an *in silico* PCR on homogeneous databases containing only sequences of the target group (details in Supplementary Material).

All primers were tested and their annealing temperatures optimized in the laboratory on DNA extracts from single specimens of each target organism group (details in Supplementary Material). Additionally, primers were tested on human and chicken DNA, as these are common laboratory contaminants (Leonard *et al.* 2007) and we aimed to exclude their amplification. PCRs were carried out in 10 µl volumes containing 1 µl of DNA, 0.5 µM of each primer, 1 mM dNTPs, 2.5 mM $MgCl_2$, 1× PCR buffer and 0.4 U Ampli*Taq* DNA Polymerase (Applied Biosystems). PCR conditions were 2 min at 94°C, followed by 55 – 60 cycles of 94°C for 30 sec, T_a (see Table 1) for 30 sec, 72°C for 30 sec, and a final extension of 10 min at 72°C.

Molecular genetic work on sediment and soil samples

The intact frozen cores were subsampled from within the cores with sterile scalpels in the ancient DNA laboratory at the Centre for Geogenetics in Copenhagen. Extraction of total DNA was carried out from 7–10 g of material. DNA was extracted using the PowerMax^{TM} Soil DNA Isolation Kit (MOBIO), with the Powerbead solution replaced by 12 ml of the following buffer: 0.96 ml C1 buffer (from PowerMax^{TM} Soil DNA Isolation Kit), 50 mM Tris/HCl, 20 mM EDTA, 150 mM NaCl, 50 mM DDT, 2 mM PTB and 0.8 mg proteinase K. Samples were digested with rotation overnight at 56°C, and the remainder of the extraction was carried out according to the manufacturer's instructions. The samples were eluted in 2 ml elution buffer.

PCRs were set up in the ancient DNA laboratory at the Natural History Museum in Oslo. For the modern soil samples, DNA was added to the PCR mix in a pre-PCR laboratory dedicated to sensitive (but not ancient) samples. PCR reactions were performed in 12.5 µl or 25 µl volumes containing 0.625 or 1.25 U Platinum® *Taq* High Fidelity DNA Polymerase (Invitrogen), 1× PCR buffer, 2 mM $MgSO_4$, 1 mM dNTPs, 0.2 µM of each primer, 0.8 mg/ml Bovine Serum Albumin (BSA) and 1–3 µl of DNA extract. Initial laboratory tests of the primers designed for Coleoptera showed that these primers also amplify human DNA when a high cycle number is used (here 55 cycles). Therefore, a blocking primer (5'–3' TTCTCGTCTTGCTGTGTCATGCC) was added to the PCR at a 10-fold concentration of the amplification primers (Vestheim & Jarman 2008). Thermal profiles were as described above for primer testing, but the extension during each cycle was performed at 68°C as recommended for the polymerase used. A subset of the positive PCRs was cloned using the Topo TA cloning kit (Invitrogen), and up to 12 clones from each cloning reaction were sequenced on an ABI 3730 sequencer.

Sequence analysis and taxonomic assignments

Clone sequences were analysed using CodonCode Aligner (version 3.6.1). To exclude errors and artefacts from being counted as true variation, clone sequences were only considered if they differed from other sequences by more than one base, or if they were present in three or more clones. If a single clone sequence displayed only a single substitution from an otherwise more common sequence, this substitution was considered likely to be an artefact and the sequence was included in a common consensus sequence.

Taxonomic assignment of the sequences was carried out using the following two approaches:

1) The best matching sequence was determined using the program ecoTag (www.grenoble.prabi.fr/trac/OBITools). This program determines identity between the query sequence and each sequence in a specified reference database through calculating the length of the longest common subsequence (LCS) by an exact algorithm corresponding to a global alignment algorithm (Ullman *et al.* 1976). Identity percent is subsequently computed by dividing the LCS length by the length of the longest sequence involved in the alignment. For each of the primer pairs a respective reference database for ecoTag was created by *in silico* PCR on the EMBL standard sequences, release 107, allowing 5 mismatches between the primer and the target sequences. To create databases with a secure taxonomy, the ecoPCR output was filtered to merge unique sequences for each taxon in the database, and to include only sequences for which complete taxonomic information is available. For the fungi primers, this filtering was not performed in order to retain database sequences obtained from uncultured organisms. With the settings used, ecoTag displays the taxon with the single closest similarity to the query sequence – either a species, if there is a single best matching sequence, or a higher-level taxon if there are multiple sequences with an equally good match. Taxonomic assignment obtained with ecoTag was only considered if the similarity was over 75%, and no assignment was

considered for sequences with a length below 18 bp (the minimum length for a sequence to be used as a specific primer; Dieffenbach *et al.* 1993).

2) As only the best matches were considered in ecoTag, we performed analyses using the Statistical Assignment Package (Munch *et al.* 2008) to obtain measures of confidence for the assignment of sequences to taxonomic groups. This program compiles a set of homologues to the query sequence using NetBlast searches against GenBank and then uses a Bayesian approach to assign a probability that a sequence belongs to a specific taxonomic group. For some sequences, taxonomic assignment was not possible, either because an insufficient number of homologues could be retrieved to proceed with the Bayesian analysis, or because the closest matching sequences were from uncultured organisms with no associated taxonomical information.

Results

Characteristics of the identified metabarcodes

The metabarcoding markers designed for each group are listed in Table 1. For fungi, the marker includes one novel primer (5.8S_fungi) used in combination with one previously published primer that is recommended by the International Fungal Barcoding Group (ITS5; White *et al.* 1990). All other primers are newly designed. The most promising metabarcoding markers for the animal groups were found to be located on the mitochondrial 12S or 16S rRNA genes, not in the standard barcoding region for animals (COI; Hebert *et al.* 2003). For birds, for which we screened complete mitochondrial genomes, the 12S fragment selected was the most optimal mitochondrial metabarcoding marker discovered. The bryophyte marker flanks the P6 loop of the *trn*L chloroplast intron, as previously selected for vascular plants (Taberlet *et al*. 2007). Our markers are hereafter referred to as Fungi_ITS (fungi), Bryo_P6 (bryophytes), Ench_12S (enchytraeids), Coleop_16S (Coleoptera) and Aves_12S (birds).

The coverage of the primers in the target organism groups calculated from *in silico* PCR was high, ranging from 86% (Bryo_P6) to 100% (Aves_12S; Table 1). The taxonomic resolution capacity was more variable among the markers: at the species level, the highest resolution is shown by Ench_12S (100%), followed by Coleop_16S (71.73%), Fungi_ITS (62.84%), Aves_12S (56.36%) and Bryo_P6 (30.57 %) (Fig. 1). The taxonomic resolution capacity increased successively from the species to the family level for bryophytes and birds, but not for fungi and Coleoptera, for which the resolution at the genus level (72.11% and 93.60%, respectively) was higher than at the family level (60.36% and 91.27%, respectively).

The median amplicon length of the metabarcodes ranged between 50 and 100 bp excluding primers, except for the Fungi_ITS amplicons, which had a median length close to 200 bp and more variation (Fig. 2; only ~2.5 % of the amplicons were more than 300 bp long).

Amplification success in recent and ancient arctic soil and sediment samples

Considerable variation in *in vitro* amplification success from soil and sediment samples was observed among the metabarcoding markers (Table 2). In the modern soils from the Varanger Peninsula, all samples amplified using the markers

Fungi_ITS, Bryo_P6 and Coleop_16S, and 67% and 13% amplified using Ench_12S and Aves_12S, respectively. Amplification success was substantially lower in the ancient samples from Duvanny Yar and Main River. The highest amplification success was achieved with the marker Fungi_ITS, which yielded positive amplifications in 50% of the ancient permafrost samples. By contrast, no positive amplification of ancient samples was achieved for Ench_12S. None of the four Main River samples older than 26 590 ± 180 yr BP gave positive results with any of the primers, but Duvanny Yar samples beyond this age could be amplified using Fungi_ITS, Bryo_P6 and Coleop_16S. Both Fungi_ITS and Bryo_P6 showed positive amplification in one of two extraction blanks, but none of the clone sequences retrieved corresponded to any of the sample sequences and thus did not compromise our results (Table S6).

Marker specificity as evaluated from in silico *and* in vitro *amplifications*

The specificity of the metabarcodes was evaluated and compared for the *in silico* and the *in vitro* amplifications. For this evaluation, the results from the modern and ancient soils were merged (Fig. 3). The *in silico* PCR results are based on the number of species amplified, while the *in vitro* PCR results are based on the number of clone sequences retrieved. Only clones with inserts that differed from primer dimers are reported (details on identification of the clones in the Supplementary Material).

For three of the five metabarcoding markers (Fungi_ITS, Bryo_P6, Aves_12S), both the *in silico* PCR and the *in vitro* PCR primarily amplified the target organism groups (Fig. 3, Table S6). However, eight out of 10 clone sequences obtained with the Aves_12S primers were identical to those of chicken (*Gallus gallus*;

Table S6), a common laboratory contaminant (Leonard *et al.* 2007). The remaining two sequences, retrieved from a Main River sample with an age of 26 590 ± 180 yr BP, were identified as passeriformes. Multiple cloning attempts of the two positive Aves_12S products from the modern soil samples (Table 2) failed to yield any sequences other than a primer multimer in one of the two samples.

For the markers Ench_12S and Coleop_16S, the target organism group did not constitute the majority of amplified sequences *in silico* (11 and 33%, respectively; Fig. 3). Nonetheless, for Ench_12S, all recovered *in vitro* sequences were inferred to stem from enchytraeids. In this case, the majority of the species amplified *in silico* were amphibians, which are not expected widely in the Arctic. With the Coleop_16S primers, only a single of the 57 clone sequences (from a Duvanny Yar sample with an age of 45 000 ± 2 000 yr BP) was inferred to stem from a beetle. Some sequences were highly divergent and identified as an enchytraeid, a reindeer (*Rangifer tarandus*) and a cow (*Bos taurus*). Notably, the single sample that amplified from Main River only yielded sequences that were identical to those of cow, a common laboratory contaminant (Leonard *et al.* 2007).

Discussion

Total soil DNA is a largely untapped resource for recent and ancient biodiversity information of eukaryotic organisms, but its potential for the study of diverse organism groups has not been comparatively assessed before. For this purpose, we here present a suite of new metabarcoding markers for ecologically highly important organism groups (bryophytes, enchytraeids, beetles and birds), and a new primer for metabarcoding of fungi. We evaluate their performance *in silico* and *in vitro* by amplification of modern and ancient arctic soil and sediment samples.

Taxonomic resolution capacity of the metabarcodes

The large differences in taxonomic resolution capacity of the metabarcodes (Fig. 1) can partly be attributed to the fact that the target groups varied considerably in size, age and evolutionary divergence among their members. The high taxonomic resolution capacity for the Ench_12S primer set (100% at the species level) exemplifies that, even with very short amplicons (~50 bp), reliable taxonomic identification is possible. This has also been demonstrated for all major eukaryotic kingdoms for fragments of the COI gene with lengths between 100 and 250 bp (Hajibabaei *et al.* 2006; Meusnier *et al.* 2008), but in this case, the suggested primer sites are not sufficiently conserved (Ficetola *et al.* 2010). The limiting factor for obtaining good metabarcodes is obviously not only the length of the sequence amplified, but the possibility to place primers specific to a target group around highly variable sequence fragments. Such conserved primers are less easily found for large groups with a high overall level of evolutionary divergence.

Among the new metabarcodes the taxonomic resolution capacity at the species level is lowest for Bryo_P6 (~30%). Nonetheless, this is ~10% higher than the resolution of the P6-loop in vascular plants (Taberlet *et al.* 2007). Like in vascular plants, the taxonomic resolution of the bryophyte P6 loop increases considerably when the set of possible sequences considered is reduced, for example, to the 400 most important arctic bryophyte species (~46% resolution to the species level, unpublished data).

The taxonomic resolution capacity of a barcoding marker calculated as the ratio of unambiguously identified taxa for a given taxonomic level (specificity index B_s, Ficetola *et al.* 2010) is expected to increase with increasing taxonomic level, such that more families than genera and species can be identified. The metabarcodes for bryophytes and birds showed this pattern, but not those for fungi and beetles (Fig. 1). This could be caused by difficulties with higher-level classifications in these two groups (Hibbett *et al.* 2007; Beutel *et al.* 2009), and/or by erroneous taxonomic assignments in GenBank. Alternatively, as these are both very large groups with a high level of evolutionary divergence, the lower resolution at the level of family could also be caused by homoplasy in the short marker sequences.

Suitability of the markers for amplifying the target organism groups

The *in silico* PCR approach offers a simple way to evaluate the performance of the primers, and we have shown here that such an evaluation provides quite a realistic prediction of the *in vitro* performance of the markers. The three primer sets designed for fungi, bryophytes and birds showed high specificity to the target groups, both in the *in silico* PCR and in the *in vitro* PCR. The other two primer sets, designed for beetles and enchytraeids, were not highly specific in the *in silico* PCR, but the *in vitro* products for Ench_12S yielded only enchytraeids. This can be explained by the fact that the other major groups potentially amplified by these primers (e.g. Amphibia, Cephalopoda) are not expected in the arctic environment.

In contrast, the *in silico* amplification of the marker Coleop_16S showed low specificity to Coleoptera (33%), but the specificity obtained from the *in vitro* experiments was even lower (2%). A number of non-arthropod sequences were retrieved, such as the putatively endogenous sequences of an enchytraeid, *Cognettia sphagnetorum,* and a mammal, *Rangifer tarandus* (reindeer), and sequences of the putative laboratory contaminant *Bos* (cow). A comparison of the closest matching sequence in GenBank with the primer sequences revealed that they displayed at most a single mismatch to any of the Coleop_16S primers – but in all cases one of these mismatches occurred at the second base on the 3' end of the primer. With the restrictions imposed in the *in silico* PCR (three mismatches allowed, but no mismatches in the last two bases at the 3' end), these amplifications were not predicted. The discrepancy between the specificity predicted by the ecoPCR results and that observed in the *in vitro* PCRs demonstrates the limitations of *in silico* PCR. Variation caused by PCR conditions (e.g. number of cycles, annealing temperature) and differences in template concentrations cannot obviously be taken into account by a pattern-matching algorithm. Hence, the amplification of taxa such as reindeer and cow is not entirely surprising when using a high number of cycles. Nonetheless, competition between templates should result in preferential amplification of coleopteran or other arthropod DNA. It is therefore notable that Coleoptera are not readily amplified from ancient sediment samples, despite the presence of Coleoptera exoskeleton remains in these environments (e.g., Sher *et al.* 2005; Elias 2006), also at the localities Main River (Kuzmina *et al.* 2011) and Duvanny Yar (Alfimov *et al.* 2003).

It should be noted that employing as many as 55 cycles in PCR can lead to the formation of chimeric sequences (Meyerhans *et al.* 1990) and other artefacts, and therefore cycle number should be minimised. However, high numbers of cycles are commonly used in ancient environmental DNA studies (e.g. Willerslev *et al.* 2003; Haile *et al.* 2009) because such samples have a mixed template pool with low initial copy numbers, causing the reactions to behave stochastically. It is however obvious that if included in further analyses, artefacts lead to erroneously increased estimates of the sequence diversity in the samples. Even if short sequences are less prone to chimera formation (Epp *et al.* 2011), retrieved sequence pools should be checked for chimeras (using approaches such as suggested by Creer *et al.* 2010; or specific program, e.g. Edgar *et al.* 2011) and other artefacts.

Potential for soil metabarcoding studies of different target groups in the Arctic

Of the groups investigated here, fungi were amplified with greatest success from permafrost sediment samples (71% and 29% success rate for Duvanny Yar and Main River, respectively; Table 2). A previous study on ancient fungal DNA preserved in permafrost successfully identified a wide diversity of fungal taxa in samples as old as 300 000 to 400 000 years (Lydolph *et al.* 2005). The generally high success for fungi is not unexpected as fungal DNA is a dominant component of total soil DNA (Pietramellara *et al.* 2009), present in a higher proportion than plant DNA or even than bacterial DNA in central European agricultural soils (Gangneux *et al.* 2011). Fungal species in the arctic ecosystem are furthermore cryoprotected by a range of different mechanisms (Ozerskaya *et al.* 2009), and it has even been suggested that microbial communities can show long-term viability in the permafrost (Lewis *et al.* 2008; Coolen *et al.* 2011). This could explain that the retrieval of good quality DNA from fungi is higher than for any of the other taxa studied here.

Amplification rates for bryophyte DNA on the other hand were high in the recent soil (100%), but this was contrasted by a low success rate for the Pleistocene samples, indicating that the amount of bryophyte DNA in the ancient sediment samples is very low. Given the ecological importance of bryophytes in the Arctic, this is surprising. However, it is possible that the bryophyte DNA content in soils is lower than that of vascular plants because of their growth form and anatomy. Most bryophyte growth and productivity occurs above ground (Lindo & Gonzalez 2010) and below ground they lack substantial somatic tissue such as true roots, which are thought to be one of the primary sources of vascular plant DNA in soil (Willerslev *et al*. 2003), particularly through the sloughing off of root cap cells (Levy-Booth *et al.* 2007). Furthermore, bryophytes contain secondary metabolites (e.g. Xie & Lou 2009), which are known to enhance DNA degradation. This could potentially cause proportionally higher DNA degradation directly after cell lysis in bryophytes compared to other organism groups.

As for bryophytes, amplification success in ancient samples was not high for the invertebrate groups. Only a single beetle sequence could be retrieved from a Pleistocene sample. Notably this sample had an age of 45 000 ± 2 000 yr BP, while no beetle DNA was found in any of the modern samples. This suggests that beetle DNA can potentially be preserved for long time periods, but the potential of sedimentary DNA for biodiversity screenings of beetles, both in modern and in ancient samples, currently seems limited. Information on beetle paleocommunities can more easily be gained by identifying exoskeleton macrofossils (e.g. Kuzmina *et al.* 2011). Unfortunately this cannot be done for enchytraeids, which often dominate soil faunal communities in the Arctic, particularly in terms of biomass (Briones *et al.* 2007), but leave no visible fossil traces. No amplification of enchytraeids was achieved from any of the Pleistocene samples, whereas for modern samples amplification success was quite high (67% in total), especially in the heath samples (87.5%). It is noteworthy that this amplification success could be obtained even from samples collected in winter, when population densities of enchytraeids are lowest (Birkemoe *et al.* 2000).

Finally, both our study and previous studies using cave sediments (Hofreiter *et al.* 2003; Haile *et al.* 2007) have retrieved putatively endogenous avian sequences from soil, although the retrieval rate is not very high. The source of bird DNA in soil could be faeces, which has been shown to contain DNA in modern populations (Regnaut *et al.* 2006; Mäki-Petäys *et al.* 2007), or dead animals. Unfortunately, our results were also confounded by the presence of contaminant chicken DNA, which will readily amplify with the avian metabarcoding primers. This problem cannot be circumvented easily, for example by the inclusion of a blocking primer (Vestheim & Jarman 2008; Gigli *et al.* 2009), because of the lack of sufficient chicken-specific mutations in the amplicon compared to other arctic Galliformes (e.g. ptarmigan). Dominant contaminant DNA may bias the PCR, hampering the retrieval of endogenous DNA present in only low concentrations (Boessenkool *et al.* online early). Therefore we cannot fully evaluate the DNA preservation and subsequent potential for diversity reconstruction of bird diversity through sedimentary ancient DNA in the Arctic.

Concluding remarks

The metabarcoding markers developed here testify that high taxonomic resolution and high specificity to target groups is achievable and can be predicted quite well using bioinformatic tools. The approach was most promising for fungi, bryophytes and enchytraeids in the recent soil, while amplification success dropped substantially in the Pleistocene samples. As no ancient samples younger than 15810 ± 75 yr BP were tested here, the potential for historical studies on shorter timescales might nonetheless be considerably larger than our results might seem to indicate.

For modern samples, metabarcoding approaches have great practical potential as an efficient and cost-effective means to conduct biodiversity screenings for ecological surveys, diet studies (Valentini *et al.* 2009a) and biological monitoring programs. Bryophyte community composition, for example, can be used as bioindicator to monitor heavy metal pollution (Denayer *et al.* 1999; Nimis *et al.* 2002), and enchytraeid diversity is suitable for ecological soil classification and assessment schemes (Jaensch *et al.* 2005). The full potential of metabarcoding can be exploited by coupling next-generation sequencing techniques with identification using reliable reference databases (e.g. Sønstebø *et al*. 2010).

Acknowledgements

This work is part of the BarFrost project supported by the Research Council of Norway (grant no. 191627/V40 to CB). Analyses were performed using the computing facilities provided by the Norwegian Metacenter for Computational Science (Notur). We thank Emilia Rota for identification of enchytraeid voucher specimens.

References

Alfimov AV, Berman DI, Sher AV (2003) Tundra-steppe insect assemblages and reconstruction of Late Pleistocene climate in the lower reaches of the Kolyma River. *Zoologichesky Zhurnal* **82**, 281-300.

Bellemain EP, Carlsen T, Brochmann C, Coissac E, Taberlet P, Kauserud H (2010) ITS as an environmental DNA barcode for fungi: an *in silico* approach reveals potential PCR biases. *Bmc Microbiology* **10**, 189.

Beutel RG, Friedrich F, Leschen RAB (2009) Charles Darwin, beetles and phylogenetics. *Naturwissenschaften* **96**, 1293-1312.

Birkemoe T, Coulson SJ, Somme L (2000) Life cycles and population dynamics of enchytraeids (Oligochaeta) from the High Arctic. *Canadian Journal of Zoology-Revue Canadienne De Zoologie* **78**, 2079-2086.

Blinnikov MS, Gaglioti BV, Walker DA, Wooller MJ, Zazula GD (2011) Pleistocene graminoid-dominated ecosystems in the Arctic. *Quaternary Science Reviews* **30**, 2906-2929.

Boessenkool S, Epp LS, Haile J, Bellemain E, Edwards ME, Coissac E, Willerslev E, Brochmann C (online early) Blocking human contaminant DNA during PCR allows amplification of rare mammal species from sedimentary ancient DNA. *Molecular Ecology*.

Briones MJI, Ineson P, Heinemeyer A (2007) Predicting potential impacts of climate change on the geographical distribution of enchytraeids: a meta-analysis approach. *Global Change Biology* **13**, 2252-2269.

Callaghan TV, Bjorn LO, Chernov Y, Chapin T, Christensen TR, Huntley B, Ims RA, Johansson M, Jolly D, Jonasson S, Matveyeva N, Panikov N, Oechel W, Shaver G, Elster J, Henttonen H, Laine K, Taulavuori K, Taulavuori E, Zockler C (2004) Biodiversity, distributions and adaptations of arctic species in the context of environmental change. *Ambio* **33**, 404-417.

Cochrane G, Akhtar R, Bonfield J, Bower L, Demiralp F, Faruque N, Gibson R, Hoad G, Hubbard T, Hunter C, Jang M, Juhos S, Leinonen R, Leonard S, Lin Q, Lopez R, Lorenc D, McWilliam H, Mukherjee G, Plaister S, Radhakrishnan R, Robinson S, Sobhany S, Hoopen PT, Vaughan R, Zalunin V, Birney E (2009) Petabyte-scale innovations at the European Nucleotide Archive. *Nucleic Acids Research* **37**, D19-D25.

Coolen MJL, Saenz JP, Giosan L, Trowbridge NY, Dimitrov P, Dimitrov D, Eglinton TI (2009) DNA and lipid molecular stratigraphic records of haptophyte succession in the Black Sea during the Holocene. *Earth and Planetary Science Letters* **284**, 610-621.

Coolen MJL, van de Giessen J, Zhu EY, Wuchter C (2011) Bioavailability of soil organic matter and microbial community dynamics upon permafrost thaw. *Environmental Microbiology* **13**, 2299-2314.

Creer S, Fonseca VG, Porazinska DL, Giblin-Davis RM, Sung W, Power DM, Packer M, Carvalho GR, Blaxter ML, Lambshead PJD, Thomas WK (2010) Ultrasequencing of the meiofaunal biosphere: practice, pitfalls and promises. *Molecular Ecology* **19**, 4-20.

De Rijk P, Robbrecht E, de Hoog S, Caers A, Van de Peer Y, De Wachter R (1999) Database on the structure of large subunit ribosomal RNA. *Nucleic Acids Research* **27**, 174-178.

Denayer FO, Van Haluwyn C, de Foucault B, Schumacker R, Colein P (1999) Use of bryological communities as a diagnostic tool of heavy metal soil contamination (Cd, Pb, Zn) in northern France. *Plant Ecology* **140**, 191-201.

Dieffenbach CW, Lowe TM, Dveksler GS (1993) General concepts for PCR primer design. *Genome Research* **3**, S30-S37.

Dunbar J, Takala S, Barns SM, Davis JA, Kuske CR (1999) Levels of bacterial community diversity in four arid soils compared by cultivation and 16S rRNA gene cloning. *Applied and Environmental Microbiology* **65**, 1662-1669.

Edgar RC, Haas BJ, Clemente JC, Quince C, Knight R (2011) UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics* **27**, 2194-2200.

Elias SA (2006) Quaternary beetle research: the state of the art. *Quaternary Science Reviews* **25**, 1731-1737.

Epp LS, Stoof-Leichsenring KR, Trauth MH, Tiedemann R (2011) Molecular profiling of diatom assemblages in tropical lake sediments using taxon-specific PCR and Denaturing High-Performance Liquid Chromatography (PCR-DHPLC). *Molecular Ecology Resources* **11**, 842-853.

Ficetola GF, Coissac E, Zundel S, Riaz T, Shehzad W, Bessiere J, Taberlet P, Pompanon F (2010) An *In silico* approach for the evaluation of DNA barcodes. *Bmc Genomics* **11**, 434.

Gangneux C, Akpa-Vinceslas M, Sauvage H, Desaire S, Houot S, Laval K (2011) Fungal, bacterial and plant dsDNA contributions to soil total DNA extracted from silty soils under different farming practices: Relationships with chloroform-labile carbon. *Soil Biology & Biochemistry* **43**, 431-437.

Gigli E, Rasmussen M, Civit S, Rosas A, de la Rasilla M, Fortea J, Gilbert MTP, Willerslev E, Lalueza-Fox C (2009) An improved PCR method for endogenous DNA retrieval in contaminated Neandertal samples based on the use of blocking primers. *Journal of Archaeological Science* **36**, 2676-2679.

Haile J, Froese DG, MacPhee RDE, Roberts RG, Arnold LJ, Reyes AV, Rasmussen M, Nielsen R, Brook BW, Robinson S, Demuro M, Gilbert MTP, Munch K, Austin JJ, Cooper A, Barnes I, Moller P, Willerslev E (2009) Ancient DNA reveals late survival of mammoth and horse in interior Alaska. *Proceedings of the National Academy of Sciences of the United States of America* **106**, 22352-22357.

Haile J, Holdaway R, Oliver K, Bunce M, Gilbert MTP, Nielsen R, Munch K, Ho SYW, Shapiro B, Willerslev E (2007) Ancient DNA chronology within sediment deposits: Are paleobiological reconstructions possible and is DNA leaching a factor? *Molecular Biology and Evolution* **24**, 982-989.

Hajibabaei M, Shokralla S, Zhou X, Singer GAC, Baird DJ (2011) Environmental Barcoding: A Next-Generation Sequencing Approach for Biomonitoring Applications Using River Benthos. *Plos One* **6**.

Hajibabaei M, Smith MA, Janzen DH, Rodriguez JJ, Whitfield JB, Hebert PDN (2006) A minimalist barcode can identify a specimen whose DNA is degraded. *Molecular Ecology Notes* **6**, 959-964.

Hebert PDN, Cywinska A, Ball SL, DeWaard JR (2003) Biological identifications through DNA barcodes. *Proceedings of the Royal Society of London Series B-Biological Sciences* **270**, 313-321.

Hibbett DS, Binder M, Bischoff JF, Blackwell M, Cannon PF, Eriksson OE, Huhndorf S, James T, Kirk PM, Lucking R, Lumbsch HT, Lutzoni F, Matheny PB, Mclaughlin DJ, Powell MJ, Redhead S, Schoch CL, Spatafora JW, Stalpers JA, Vilgalys R, Aime MC, Aptroot A, Bauer R, Begerow D, Benny GL, Castlebury LA, Crous PW, Dai YC, Gams W, Geiser DM, Griffith GW, Gueidan C, Hawksworth DL, Hestmark G, Hosaka K, Humber RA, Hyde KD, Ironside JE, Koljalg U, Kurtzman CP, Larsson KH, Lichtwardt R, Longcore J, Miadlikowska J, Miller A, Moncalvo JM, Mozley-Standridge S, Oberwinkler F, Parmasto E, Reeb V, Rogers JD, Roux C, Ryvarden L, Sampaio JP, Schussler A, Sugiyama J, Thorn RG, Tibell L, Untereiner WA, Walker C, Wang Z, Weir A, Weiss M, White MM, Winka K, Yao YJ, Zhang N (2007) A higher-level phylogenetic classification of the Fungi. *Mycological Research* **111**, 509-547.

Hickson RE, Simon C, Copper A, Spicer GS, Sullivan J, Penny D (1996) Conserved sequence motifs, alignment, and secondary structure for the third domain of animal 12S rRNA. *Molecular Biology and Evolution* **13**, 150-169.

Hofreiter M, Mead JI, Martin P, Poinar HN (2003) Molecular caving. *Current Biology* **13**, R693-R695.

Jaensch S, Rombke J, Didden W (2005) The use of enchytraeids in ecological soil classification and assessment concepts. *Ecotoxicology and Environmental Safety* **62**, 266-277.

Kalendar R, Lee D, Schulman AH (2009) FastPCR software for PCR primer and probe design and repeat search. *Genes, Genomes and Genomics,* **3**, 1-14.

Kuzmina SA, Sher AV, Edwards ME, Haile J, Yan EV, Kotov AV, Willerslev E (2011) The late Pleistocene environment of the Eastern West Beringia based on the principal section at the Main River, Chukotka. *Quaternary Science Reviews* **30**, 2091-2106.

Leonard JA, Shanks O, Hofreiter M, Kreuz E, Hodges L, Ream W, Wayne RK, Fleischer RC (2007) Animal DNA in PCR reagents plagues ancient DNA research. *Journal of Archaeological Science* **34**, 1361-1366.

Levy-Booth DJ, Campbell RG, Gulden RH, Hart MM, Powell JR, Klironomos JN, Pauls KP, Swanton CJ, Trevors JT, Dunfield KE (2007) Cycling of extracellular DNA in the soil environment. *Soil Biology & Biochemistry* **39**, 2977-2991.

Lewis K, Epstein S, Godoy VG, Hong SH (2008) Intact DNA in ancient permafrost. *Trends in Microbiology* **16**, 92-94.

Lindo Z, Gonzalez A (2010) The Bryosphere: an integral and influential component of the earth's biosphere. *Ecosystems* **13**, 612-627.

Lydolph MC, Jacobsen J, Arctander P, Gilbert MTP, Gilichinsky DA, Hansen AJ, Willerslev E, Lange L (2005) Beringian paleoecology inferred from permafrost-preserved fungal DNA. *Applied and Environmental Microbiology* **71**, 1012-1017.

Mäki-Petäys H, Corander J, Aalto J, Liukkonen T, Helle P, Orell M (2007) No genetic evidence of sex-biased dispersal in a lekking bird, the capercaillie (Tetrao urogallus). *Journal of Evolutionary Biology* **20**, 865-873.

Meusnier I, Singer GAC, Landry JF, Hickey DA, Hebert PDN, Hajibabaei M (2008) A universal DNA mini-barcode for biodiversity analysis. *Bmc Genomics* **9**, 214.

Meyerhans A, Vartanian JP, Wainhobson S (1990) DNA recombination during PCR. *Nucleic Acids Research* **18**, 1687-1691.

Munch K, Boomsma W, Huelsenbeck JP, Willerslev E, Nielsen R (2008) Statistical assignment of DNA sequences using Bayesian phylogenetics. *Systematic Biology* **57**, 750-757.

Nimis PL, Fumagalli F, Bizzotto A, Codogno M, Skert N (2002) Bryophytes as indicators of trace metals pollution in the River Brenta (NE Italy). *Science of the Total Environment* **286**, 233-242.

Ozerskaya S, Kochkina G, Ivanushkina N, Gilichinsky DA (2009) Fungi in permafrost. *Soil Biology* **16**, 85-95.

Pääbo S, Poinar H, Serre D, Jaenicke-Despres V, Hebler J, Rohland N, Kuch M, Krause J, Vigilant L, Hofreiter M (2004) Genetic analyses from ancient DNA. *Annual Review of Genetics* **38**, 645-679.

Pietramellara G, Ascher J, Borgogni F, Ceccherini MT, Guerri G, Nannipieri P (2009) Extracellular DNA in soil and sediment: fate and ecological relevance. *Biology and Fertility of Soils* **45**, 219-235.

Pompanon F, Coissac E, Taberlet P (2011) Metabarcoding, a new way to analyze biodiversity. *Biofutur* **30**, 30-32.

Rappé MS, Giovannoni SJ (2003) The uncultured microbial majority. *Annual Review of Microbiology* **57**, 369-394.

Raupach MJ, Astrin JJ, Hannig K, Peters MK, Stoeckle MY, Wagele JW (2010) Molecular species identification of Central European ground beetles (Coleoptera: Carabidae) using nuclear rDNA expansion segments and DNA barcodes. *Frontiers in Zoology* **7**.

Ravolainen VT, Bråthen KA, Ims RA, Yoccoz NG, Henden J-A, Killengreen ST (2011) Rapid, landscape scale responses in riparian tundra vegetation to exclusion of small and large mammalian herbivores. *Basic and Applied Ecology* **12**, 643-653.

Regnaut S, Lucas FS, Fumagalli L (2006) DNA degradation in avian faecal samples and feasibility of non-invasive genetic studies of threatened capercaillie populations. *Conservation Genetics* **7**, 449-453.

Riaz T, Shehzad W, Viari A, Pompanon F, Taberlet P, Coissac E (2011) ecoPrimers: inference of new DNA barcode markers from whole genome sequence analysis. *Nucleic Acids Research* **1**, gkr732.

Sher AV, Kuzmina SA, Kuznetsova TV, Sulerzhitsky LD (2005) New insights into the Weichselian environment and climate of the East Siberian Arctic, derived from fossil insects, plants, and mammals. *Quaternary Science Reviews* **24**, 533-569.

Shokralla S, Zhou X, Janzen DH, Hallwachs W, Landry JF, Jacobus LM, Hajibabaei M (2011) Pyrosequencing for Mini-Barcoding of Fresh and Old Museum Specimens. *Plos One* **6**.

Sonnenberg R, Nolte AW, Tautz D (2007) An evaluation of LSU rDNA D1-D2 sequences for their use in species identification. *Frontiers in Zoology* **4**, 6.

Sønstebø JH, Gielly L, Brysting AK, Elven R, Edwards M, Haile J, Willerslev E, Coissac E, Rioux D, Sannier J, Taberlet P, Brochmann C (2010) Using next-generation sequencing for molecular reconstruction of past Arctic vegetation and climate. *Molecular Ecology Resources* **10**, 1009-1018.

Stackebrandt E, Liesack W, Goebel BM (1993) Bacterial diversity in a soil sample from a subtropical australian environment as determined by 16S rDNA analysis. *Faseb Journal* **7**, 232-236.

Taberlet P, Coissac E, Pompanon F, Gielly L, Miquel C, Valentini A, Vermat T, Corthier G, Brochmann C, Willerslev E (2007) Power and limitations of the chloroplast trnL (UAA) intron for plant DNA barcoding. *Nucleic Acids Research* **35**, e14.

Ullman JD, Aho AV, Hirschberg DS (1976) Bounds on the complexity of the longest common subsequence problem. *Journal of the ACM* **23**, 1-12.

Valentini A, Miquel C, Nawaz MA, Bellemain E, Coissac E, Pompanon F, Gielly L, Cruaud C, Nascetti G, Wincker P, Swenson JE, Taberlet P (2009a) New perspectives in diet analysis based on DNA barcoding and parallel pyrosequencing: the trnL approach. *Molecular Ecology Resources* **9**, 51-60.

Valentini A, Pompanon F, Taberlet P (2009b) DNA barcoding for ecologists. *Trends in Ecology & Evolution* **24**, 110-117.

Vestheim H, Jarman SN (2008) Blocking primers to enhance PCR amplification of rare sequences in mixed samples - a case study on prey DNA in Antarctic krill stomachs. *Frontiers in Zoology* **5**, 12.

White T, Bruns T, Lee S, Taylor J (1990) Amplification and direct sequencing of fungal ribosomal RNA genes for phylogenetics. In: *PCR-protocols A guide to methods and applications.* (eds MA Innis, DH Gelfand, JJ Sninski, TJ White), pp. 315-322. Academic Press, San Diego.

Willerslev E, Cappellini E, Boomsma W, Nielsen R, Hebsgaard MB, Brand TB, Hofreiter M, Bunce M, Poinar HN, Dahl-Jensen D, Johnsen S, Steffensen JP, Bennike O, Schwenninger JL, Nathan R, Armitage S, de Hoog CJ, Alfimov V, Christl M, Beer J, Muscheler R, Barker J, Sharp M, Penkman KEH, Haile J, Taberlet P, Gilbert MTP, Casoli A, Campani E, Collins MJ (2007) Ancient biomolecules from deep ice cores reveal a forested Southern Greenland. *Science* **317**, 111-114.

Willerslev E, Hansen AJ, Binladen J, Brand TB, Gilbert MTP, Shapiro B, Bunce M, Wiuf C, Gilichinsky DA, Cooper A (2003) Diverse plant and animal genetic records from Holocene and Pleistocene sediments. *Science* **300**, 791-795.

Xie CF, Lou HX (2009) Secondary Metabolites in Bryophytes: An Ecological Aspect. *Chemistry & Biodiversity* **6**, 303-312.

Data accessibility

All environmental clone sequences are deposited in the Dryad database: doi... .

Figures

Fig. 1 Taxonomic resolution capacity (specificity index B_s, Ficetola *et al.* 2010) of the metabarcoding markers Fungi_ITS, Bryo_P6, Ench_12S, Coleop_16S, Aves_12S within the respective target taxa. This was calculated from the output obtained using *in silico* PCR on all standard sequences in the EMBL database, release 107.

Fig. 2 Boxplots of the amplicon length variation in the metabarcoding markers Fungi_ITS, Bryo_P6, Ench_12S, Coleop_16S, Aves_12S. This was determined from the output obtained using *in silico* PCR on all standard sequences in the EMBL database, release 107. Length is given excluding primer sequences, and outliers are not shown.

Fig. 3 Comparison of the *in silico* PCR (left) and *in vitro* PCR and cloning of arctic soil and sediment samples (right) using the different markers designed in this study. Actual numbers (N) and percentages (%) of species and clone sequences retrieved from the *in silico* PCRs and from the *in vitro* PCRs are given. The *in silico* PCR was performed on all standard sequences in the EMBL database, release 107. Results obtained from the recent and ancient arctic samples were merged.

Table 1. Primer characteristics of the metabarcoding markers. T_a = annealing temperature based on optimisation in the laboratory. Taxonomic coverage is calculated as % amplified target species of the total number of target species in the database, using *in silico* PCR.

Taxon	Primer name	Primer sequence (5'-3')	Genomic region	T_a (°C)	Taxonomic coverage (%)
Fungi	ITS5	GGAAGTAAAAGTCGTAACAAGG	ITS1	55	95.2
	5.8S_fungi	CAAGAGATCCGTTGTTGAAAGTT			
Bryophytes	bryo_P6F	GATTCAGGGAAACTTAGGTTG	*trn*L P6-loop	51	86.0
	bryo_P6R	CCATTGAGTCTCTGCACC			
Enchytraeidae	Ench_12Sa	GCTGCACTTTGACTTGAC	12S	56	98.0
	Ench_12Sc	AGCCTGTGTACTGCTGTC			
Coleoptera	Coleop_16Sc	TGCAAAGGTAGCATAATMATTAG	16S	55	98.5
	Coleop_16Sd	TCCATAGGGTCTTCTCGTC			
Aves	Aves_12Sa	GATTAGATACCCCACTATGC	12S	58	100
	Aves_12Sc	GTTTTAAGCGTTTGTGCTCG			

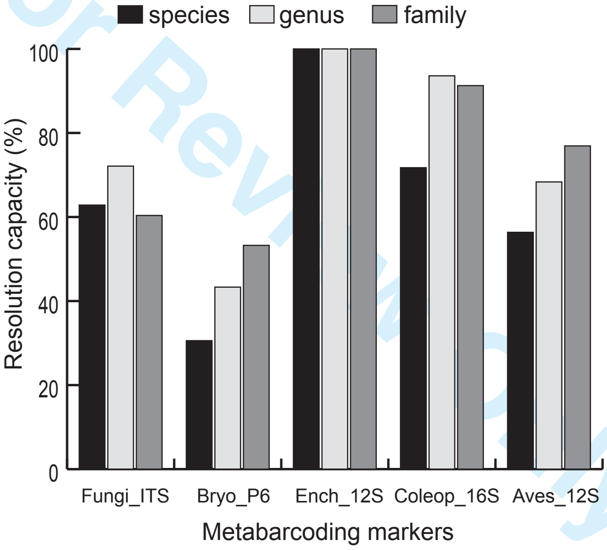
Table 2: Amplification success on soil and sediment samples for the different markers. + indicates that a positive amplification was obtained, – indicates no amplification. Sample ages, when available, are given with their laboratory identifier in uncalibrated ^{14}C yr BP and counting error.

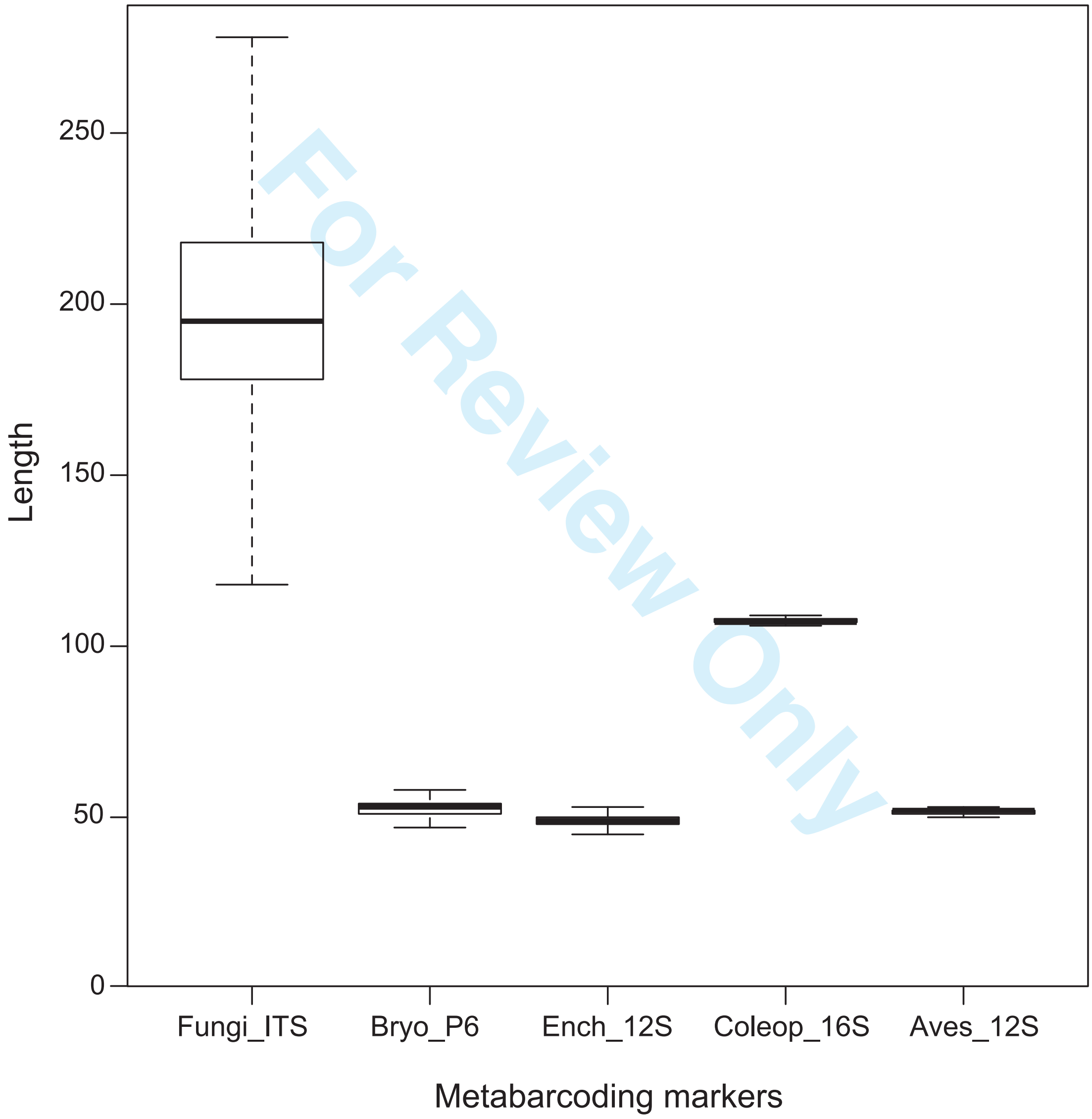
Varanger Field Sample	Sample age (Lab. identifier)	Fungi_ITS	Bryo_P6	Coleop_16S	Ench_12S	Aves_12S
ENG_A_1.2	undated	+	+	+	–	+
ENG_A_2.2	undated	+	+	+	–	–
ENG_B_1.2	modern (Poz-26576)	+	+	+	+	+
ENG_C_1.2	undated	+	+	+	+	–
ENG_D_1.2	undated	+	+	+	–	–
ENG_D_1.3	undated	+	+	+	–	–
ENG_D_2.1	undated	+	+	+	+	–
HEI_A_1.2	undated	+	+	+	+	–
HEI_A_2.2	undated	+	+	+	+	–
HEI_B_1.2	undated	+	+	+	+	–
HEI_C_1.2	undated	+	+	+	+	–
HEI_D_1.2	undated	+	+	+	–	–
HEI_D_1.3	undated	+	+	+	+	–
HEI_D_2.1	modern (Poz-26591)	+	+	+	+	–
HEI_D_2.2	undated	+	+	+	+	–
	Success rate (%):	100	100	100	67	13

Table 2 (cont.)

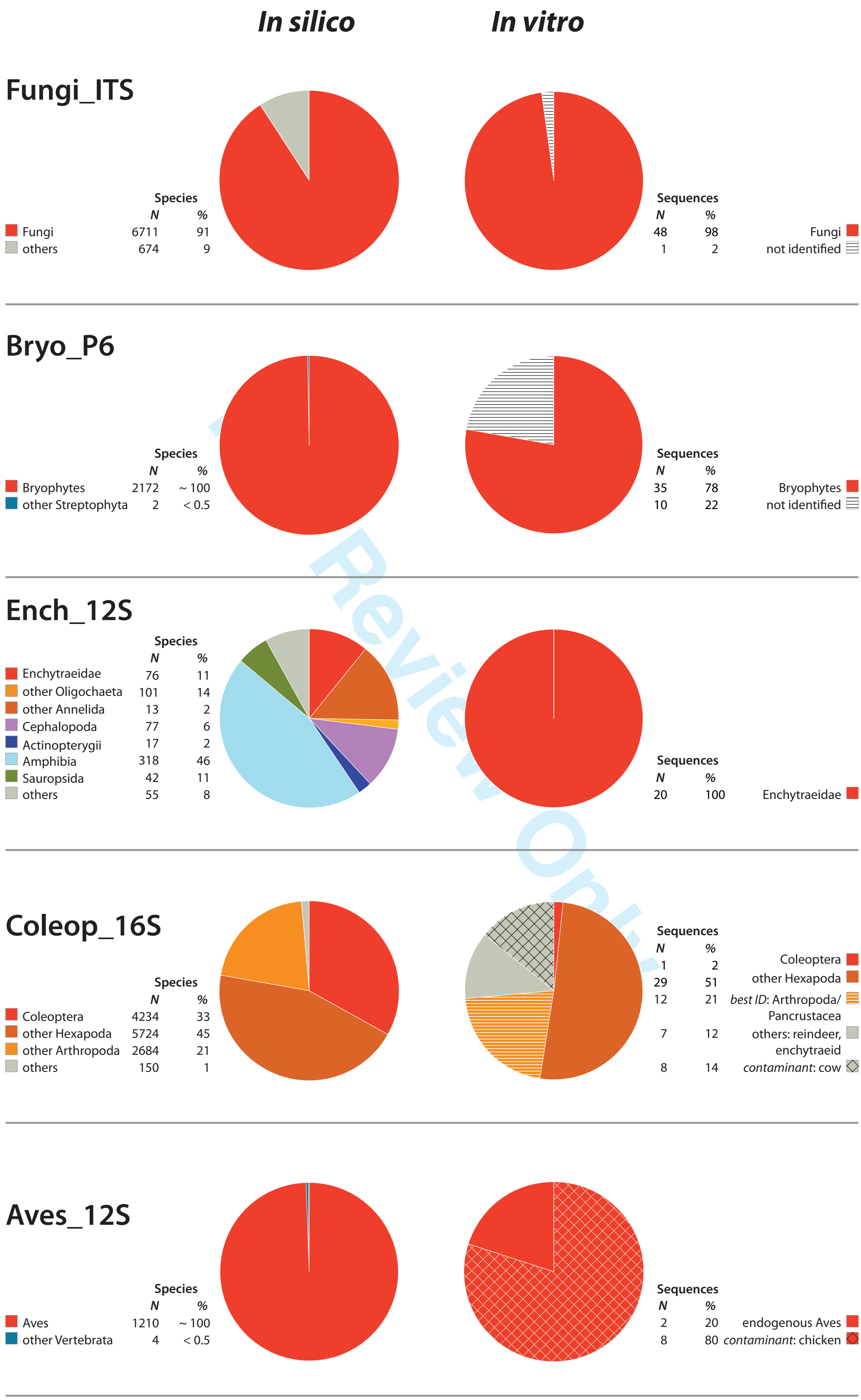
Duvanny Yar Field Sample	Sample age	Lab. identifier	Fungi_ITS	Bryo _P6	Coleop_16S	Ench_12S	Aves_12S
69	16850 ± 100 yr BP	Poz-32563	–	–	–	–	–
57A	19780 ± 130 yr BP	Poz-32457	+	+	–	–	–
59	20670 ± 120 yr BP	Poz-32490	+	–	–	–	–
62	22900 ± 170 yr BP	Poz-32557	–	–	–	–	–
64	23630 ± 190 yr BP	Poz-32559	+	–	+	–	–
67	25340 ± 220 yr BP	Poz-32562	+	–	+	–	–
118	25830 ± 630 yr BP	Poz-32681	+	+	+	–	–
122	28530 ± 800 yr BP	Poz-32682	+	–	–	–	–
131	29900 ± 300 yr BP	Poz-32791	–	–	–	–	–
21A	>48000 yr BP	Poz-32301	+	–	+	–	–
24A	50000 ± 2000 yr BP	Poz-32232	–	–	+	–	–
28A	>45000 yr BP	Poz-32370	+	+	–	–	–
32A	>45000 yr BP	Poz-32374	+	–	–	–	–
26B	45000 ± 2000 yr BP	Poz-32368	+	–	+	–	–
		Success rate (%):	71	21	50	0	0

Main River Field Sample	Sample age	Lab. identifier	Fungi_ITS	Bryo P6	Coleop_16S	Ench_12S	Aves_12S
58A	15810 ± 75 yr BP	OxA-14930	–	–	–	–	–
55B	20030 ± 110 yr BP	Poz-28724	+	–	–	–	–
54B	20160 ± 110 yr BP	Poz-28723	+	–	–	–	–
47A	20830 ± 90 yr BP	OxA-15667	–	+	–	–	–
56B	20900 ± 110 yr BP	OxA-14958	+	–	+	–	–
39A	23210 ± 130 yr BP	OxA-15348	–	+	–	–	–
34B	23880 ± 140 yr BP	Poz-28680	+	–	–	–	+
36A	25440 ±130 yr BP	OxA-14957	–	–	–	–	–
35B	25450 ± 160 yr BP	Poz-28681	–	+	–	–	–
37B	26590 ± 180 yr BP	Poz-28682	–	–	–	–	+
33A	28190 ± 160 yr BP	OxA-15349	–	–	–	–	–
28A	29780 ±210 yr BP	OxA-14928	–	–	–	–	–
27A	30900 ± 400 yr BP	Poz-28653	–	–	–	–	–
19B	>47000 yr BP	Poz-28618	–	–	–	–	–
		Success rate (%):	29	21	7	0	14


species
genus
family
Resolution capacity (%)
0
20
40
60
80
100
Fungi_ITS
Bryo_P6
Ench_12S
Coleop_16S
Aves_12S
Metabarcoding markers



Length
0
50
100
150
200
250
Fungi_ITS
Bryo_P6
Ench_12S
Coleop_16S
Aves_12S
Metabarcoding markers


In silico
In vitro
Fungi_ITS
Species
N %
🟥 Fungi 6711 91
🟩 others 674 9
Sequences
N %
48 98 Fungi 🟥
1 2 not identified ▤
Bryo_P6
Species
N %
🟥 Bryophytes 2172 ~ 100
🟦 other Streptophyta 2 < 0.5
Sequences
N %
35 78 Bryophytes 🟥
10 22 not identified ▤
Ench_12S
Species
N %
🟥 Enchytraeidae 76 11
🟧 other Oligochaeta 101 14
🟧 other Annelida 13 2
🟪 Cephalopoda 77 6
🟦 Actinopterygii 17 2
🟦 Amphibia 318 46
🟩 Sauropsida 42 11
🟩 others 55 8
Sequences
N %
20 100 Enchytraeidae 🟥
Coleop_16S
Species
N %
🟥 Coleoptera 4234 33
🟧 other Hexapoda 5724 45
🟧 other Arthropoda 2684 21
🟩 others 150 1
Sequences
N %
1 2 Coleoptera 🟥
29 51 other Hexapoda 🟧
12 21 best ID: Arthropoda/ Pancrustacea ▤
7 12 others: reindeer, enchytraeid 🟩
8 14 contaminant: cow ▨
Aves_12S
Species
N %
🟥 Aves 1210 ~ 100
🟦 other Vertebrata 4 < 0.5
Sequences
N %
2 20 endogenous Aves 🟥
8 80 contaminant: chicken ▨