Analysing ancient DNA

Raúl J. Cano

Much of what we know about extinct organisms comes from traits that are not preserved in the fossil record. Until recently, morphological analysis was the only tool available for scientists to determine relationships for extinct fossil organisms. We now know that ‘ancient’ DNA can be preserved in the remains of extinct organisms. By targeting specific gene sequences, it may be possible to deduce biochemical characteristics and through sequence comparisons, to estimate the extent of evolutionary divergence. By comparing the amount and type of these changes, one could estimate how quickly some DNA ‘evolves’ relative to other segments, or which genes have the most flexibility or are more conserved over time. The compilation of these data would yield greater understanding of the physiology of extinct organisms and provide a much clearer picture of genetic change over time, and the mechanics behind ‘evolution’.

The isolation and characterization of fossil DNA, until recently [1], was considered unattainable as the methodologies for extracting minute quantities of partially degraded DNA and their subsequent enzymatic amplification were not available. With the advent of the polymerase chain reaction [2], a new analytical tool became available for the molecular study of fossils. It is now possible to conduct molecular studies of extinct organisms utilizing their DNA to unravel biological and evolutionary questions.

There is already a body of scientific evidence built which supports the use of DNA from extinct animals and plants for phylogenetic studies. Higuchi and Wilson [1] demonstrated that remains of a mammoth and the extinct species, the guagga, contained fragments of the original DNA. Mulli [3] reported the extraction of clonable DNA from a 2400-year-old mummy of a child. Subsequent DNA analysis revealed fragments measuring approximately 3.4 kilobase pairs (Kbp). Thomas et al. [4] isolated DNA from hair found in century-old untanned hide and a piece of dried muscle collected from an extinct marsupial wolf. This DNA was later enzymatically amplified by polymerase chain reaction (PCR) and phylogenetic studies were made. More recently, Golenberg et al. [5] isolated and analysed Magnolia chloroplast DNA from a Miocene Clarkia deposit dated 17–20 million years old. Cano et al. [6] isolated and characterized DNA from the extinct bee Proplebeia dominicana in 25–40 million-year-old Dominican amber. DeSalle et al. [7] employed DNA extracted from fossil termites to resolve phylogenetic relationships between the termites, cockroaches and mantids. Cano et al. [8] extracted DNA from a 120–135 million-year-old nemonychid weevil Lecontella pinicola. Poinar et al. [9] used DNA sequences from the extinct legume Hymenaea protera in Dominican amber in a biogeographical study in which they showed that the extinct H. protera was most closely related to the extant nemonychid weevil Lecontella pinicola. Poinar et al. [10] used DNA sequences from 25–40 million-year-old Bacillus spp. in Dominican amber inclusions to study a symbiotic relationship between Bacillus and the now extinct stingless bee Proplebeia dominicana (Figure 1). The value of fossil evidence is that it may demonstrate the condition of taxa before evolutionary divergence obscured phylogenetic relationships [11]. Because they are older, ancient fossil DNA sequences should be less divergent than extant sequences and should, therefore, have value for relating more derived extant taxa. When compared with extant DNA, ancient DNA sequences may also provide an insight into the pattern of molecular evolutionary change through time. Fossil DNA has been used to answer evolutionary questions among organisms [7], detect the presence of pathogens in museum specimens [12], study the origin of Pacific Islander populations [13], and study spatial and temporal distribution of populations [14].

An interesting question that can be addressed with fossil DNA is the ‘molecular clock hypothesis’. Fossil DNA sequence data can be used for estimating the rate and pattern of molecular change through time [15,16]. To study this pattern, it might be possible to compare typical pairwise distances derived from nucleotide sequence data measured between extant genera with the distances measured between the fossils and unrelated extant taxa.

Strategies of analysis

The explosion in the field of biotechnology has made areas of study available to molecular palaeontology that were never before possible. In terms of the analysis of DNA, the single most important technology is the polymerase chain reaction. This exponential amplification produces enough copies of the target strand of DNA to be manipulated and analysed through standard molecular techniques, such as cloning and enzymatically directed sequencing. Coupled with new and refined techniques for extraction of biomolecules tightly adhered to matrices, this technology has become a powerful tool for analysis in molecular palaeontology.

Analytical software is available, for example, CLUSTAL [17], FASTA [18], and GDE (S. Smith, pers. commun., 1994), which allows the sequences obtained by the above methods to be matched against homologous sequences from other species which have been entered into a data bank. Statistical analyses can then be performed and estimates of relatedness and genetic distance can be obtained. Phylogenetic trees based on sequence data can be constructed, using software packages such as MEGA [19], PAUP [20], and PHYLIP [21]. This allows for the objective placement of an organism within the framework of known taxa (Figure 2). It also allows any modern
Sequences can be identified. Regions of sequences. When selected regions of genes sequence for amplification and/or analysis segment from fossils, it is best to select representatives of the closest living relatives (that is, closest to the root) to modern isolates of B. sphaericus (B.SPHAER, BACRRNAGA-BACRRNAGG). Selections of gene sequences for analysis for these taxa are compared, homologous sequences increase as the DNA becomes damaged and degraded, the resulting fragment length becomes smaller.

When working with DNA putatively obtained from fossils, the selection of gene sequences for amplification and/or analysis is a crucial step. In the case of extinct organisms, for which there is no direct living representative, the genomes of the closest living relatives (based on morphological analysis) are examined for conserved sequences. When selected regions of genes for these taxa are compared, homologous sequences can be identified. Regions of homology, where at least 15 bases are identical between the two groups, are good places to start when designing primer molecules.

The size of the amplified target sequence (amplicon) is also of importance. Generally, when designing primers to amplify DNA segments from fossils, it is best to think small. The chances for successful amplifications of fossil sequences increase as the size of the amplicon decreases. As a general rule it is recommended that the selected primer pair amplify a region of the desired gene to measure ≤ 200 bp. As the fossil DNA becomes damaged and degraded, the resulting fragment length becomes smaller. Thus, amplification of small DNA segments will be more successful than that of larger segments. Table 1 illustrates the results of a study conducted in our laboratory aimed at demonstrating the reproducibility of DNA extraction from amber inclusions (dated 25–40 million years old) of the extinct bee Proplebeia dominicana, the extant bee Plebeia frontalis, and their corresponding Bacillus symbionts.

The chances for successful amplification are increased if the target gene sequence is present in multiple copies within each cell. Nuclear DNA sequences of ribosomal constituents, such as 18s and 28s rDNA, are often used in such studies. Mitochondrial DNA sequences are also good candidates because, not only are there thousands of mitochondria per cell, but the complete mitochondrial genome for many taxa have been sequenced and entered into data banks, and are available for comparative studies. For phylogenetic significance, it is desirable that the selected homologous primer sequences for known taxa flank regions of relatively high variability. This allows for better definition of phylogenetic placement than if there are relatively few changes across a broad range of taxa. Also, it is easier to tell if there is contamination with modern DNA. If DNA from ancient samples can be obtained and amplified, then analysis can reveal if any base pair changes in the sequences from ancient materials are intermediate between the modern taxa being used for comparison.

### Preservation potential of biomolecules

It is a commonly held belief, based on experimental evidence as well as extrapolated predictions based on studies of DNA...
in aqueous solution, that nucleic acids do not survive in fossil remains on a geological time-scale [4]. These assumptions, however, are being challenged by researchers who are continually pushing back the age for identification and recovery of DNA and proteins obtained from fossils preserved under rare and specific conditions [1,3,6-8, 10,13, 23-25].

The double-stranded, helical structure of DNA is more resistant to damage than single-stranded RNA [4], but its structure and chemistry make it susceptible to certain types of damage over time. Conversion of bases through hydrolytic deamination (guanine changes to xanthine, cytosine to uracil or its derivatives) and depurination (removal of the bases guanine and adenine from the sugar-phosphate backbone) affect the informational content of the molecule. Exposure to oxygen free radicals or UV radiation also damages DNA strands [26]. Mechanisms have evolved in living organisms for repairing such DNA damage as it occurs, maintaining genetic information, and preventing accumulation or errors [27]. With the death of the organism, this self-repair process stops, while enzymatic attack and exposure to water, oxygen and ultraviolet radiation continue with advancing decay. There are rare cases, however, where DNA is protected from such damage.

Exposure to water is probably the single most destructive force acting on the DNA molecule. Water has been shown to initiate strand breaks by attacking the base-sugar bonds. Where the base is lost, the chain is weakened, and eventually cleaved [26,28]. Given these facts, a crucial step in the preservation of DNA is relatively rapid dehydration of tissues. One way that this occurs is through entrapment of organisms in amber-forming resins (Figure 3).

Amber is an amorphous polymeric glass, with mechanical, dielectric and thermal features common to synthetic polymers [29]. It originates from the resin of woody plants, and is commonly recognized as sticky, odoriferous 'pitch'. Natural resins are complex mixtures of terpenoid compounds, acids, alcohols, and saccharides secreted from parenchymal cells, some of which have preservative and antimicrobial properties [30]. Resins are not restricted to the conifers but occur in a wide range of flowering plants [30]. Through the ageing processes of oxidation and polymerization, the resin becomes harder and ultimately forms the gemstone known as amber. The preservative properties of amber make it a suitable source of tissue with extractable DNA, from which genetic studies can be conducted [5-10].

What makes amber such a good preservative of DNA? Studies conducted on the trunk resin of the tree Agathis australis may provide part of the answer. First, the sugars arabinose, galactose and sucrose are present in such resins. High concentrations of these sugars in the resin would make the resin hyperosmotic to the cell, drawing water out
and achieving tissue dehydration. Under water-free conditions, biochemical reactions, including those involved in the degradation of nucleic acids and proteins, are inhibited. Microbial activity which results in the degradation of cellular components is also halted, as there is not sufficient water to carry out microbial metabolism. Secondly, alcohols such as fenethyl and communol and terpenes such as alpha-pinene, limonene and dipentene may act as fixatives to preserve tissue. Evidence of such preservative properties can be seen in the electron photomicrographs in Figure 4, which show evidence of chromatin, endoplasmic reticulum, and mitochondria of a 40 Ma midge fly in Mexican amber. Additionally, one of the oxygenated derivatives of terpene hydrocarbons is aldehyde, which may also serve as a fixative of embedded tissue.

Effective dehydration can also occur with the removal of DNA from solution. This process occurs through adsorption of DNA onto mineral surfaces. Hydroxyapatite is known to have a very strong binding affinity for DNA [31] and this component is, of course, the mineral which predominates in bone. Removal from solution through adsorption protects the molecule from attack by hydronium ions.

Another consideration in the long-term preservation of DNA is the pH of the environment. Acidic environments may increase the rate of degradation of this molecule as H⁺ ions can attack the OH groups of the sugars and the nitrogenous bases, contributing to strand breakage. Bone also sets up an alkaline environment (hydroxyapatite is a basic compound), which can favour the preservation of DNA [31-35]. However, Lindahl [28] claims that in the vicinity of 7.4, variations in pH do not seem to be a major factor in the degradation of DNA.

Oxidation is another source of DNA damage, and removing DNA from water as in amber or bone does not protect the molecule from oxidative attack. Oxygen, in its molecular state, does not attack DNA, but rather it is the formation of oxygen free radicals that attack the nitrogenous bases. Oxidative attack would be rapid at first, but then would level off [3]. It is proposed that chelation of copper or other metal ions [26] enhances the preservation of this molecule by contributing to a reducing environment, and compensating for the production of oxygen free radicals.

Exposure to ultraviolet light also causes extensive damage and degradation of DNA, and rapid burial of an organism is important to minimize the consequences of UV damage to DNA. Rapid burial is implied in the preservation of fossils such as fossil bones. It is assumed that predation, bloating, bacterial decay, scavenging, and other taphonomic processes seen today were equally active during prehistoric times in the breakdown of organic remains. To avoid total disintegration of remains by these forces, burial must have occurred relatively soon after death. This is particularly true when skeletons are found fully articulated. The assumption is made that burial occurred before the soft tissues like ligaments, muscles and skin, which hold the bones together, had undergone complete decay.

Problems of working with 'fossil' biomolecules

The extreme sensitivity of PCR, which opens the door to the direct analysis of DNA obtained from ancient materials, also poses the most complications. The fact that PCR technology can amplify as little as one molecule of DNA means that minute amounts of contaminating DNA from modern sources, such as bacteria, soil fungi, or human skin cells, can also be amplified. Indeed, any such modern contaminant would probably be amplified preferentially over ancient target molecules, owing to the probable state of degradation of the latter. It is for this reason that the selection of primer molecules used in amplification is such a crucial step, as careful design can decrease or eliminate spurious amplification of contaminating DNA. Through studies of published sequences of extant species, it is desirable to build primer molecules from regions that would prevent the amplification of DNA from the most common sources of contamination. Also, it becomes very important to run several environmental controls at each step of the isolation and amplification process. If the gene sequences chosen for amplification flank regions of variability, or regions containing insertion or deletions, then analysis of sequence data obtained from PCR amplification of ancient targets makes contamination by modern DNA much easier to detect.

Limiting access of technicians to ancient-DNA laboratories and equipment reduces potential sources for contamination. Frequent washing of laboratory surfaces with a 10 per cent bleach solution, and continual exposure of surfaces and reagents to

Figure 4: (a) Endospore of Bacillus sp. from abdominal tissue of stingless bee in Mexican amber (19-25 myo). (b) Electron photomicrograph of tissue from 40 myo fly in Baltic amber showing smooth endoplasmic reticulum. (c) Electron photomicrograph of tissue from 40 myo fly in Baltic amber showing tracheole, mitochondrion, and muscle tissue. (Courtesy of R. Hess-Poinar and G.O. Poinar Jr.)
UV light when not in use also reduces the potential for contamination, as UV light is known to cross-link DNA strands, thus making them unavailable for amplification by PCR. Keeping laboratories used in ancient-DNA work separated from any used in modern analyses is another important requirement. Likewise, separating areas for extraction of DNA from areas designated for setting up PCR reactions also minimizes the possibility of contamination.

Ultimately, however, the proof of the authenticity of any DNA presumably obtained from ancient materials comes from careful analysis of sequence data. If phylogenetic analysis of the sequences does not agree with predicted relationships based on morphological data, particularly with species such as dinosaurs which leave no modern representatives, then the DNA data must be carefully re-evaluated. Also, the analysis of at least two different genes or gene regions should be done, and the results of both should show similar or identical phylogenies, before any claims can be made regarding the sources of the DNA.

The polymerase chain reaction assay

Once the DNA from fossils has been successfully extracted [36,37], it is now ready for enzymatic amplification. Needless to say, gene selection and primer design are of primary importance and will depend upon the goals of the amplification assay. As each target DNA and its corresponding primer pair(s) are unique, the reaction and conditions and thermal cycling protocol will vary with each sequence and therefore the assay must be optimized each time a new primer set is used. The Stoffel fragment of Taq polymerase is sometimes used for initial studies of fossil DNA as this enzyme is more tolerant to fluctuation in Mg2+ concentrations and, therefore, would increase the chances for initial success.

Many fossil samples have tannins, porphyrins, hematin, and other inhibitors of the PCR reaction. For this reason, bovine serum albumin (BSA fraction V, Sigma) in the reagent mixture at concentrations of 2 μg/ml is added to the reaction mixture to palliate the inhibitory activity of fossil DNA contaminants.

Also, to reduce spurious hybridization of primers to non-homologous target DNA sequences some modification of a 'hot start' PCR should be used. We describe a method that has been largely successful in our laboratory and does not require the separate addition of polymerase to each tube or the use of wax beads. In essence, the reaction mixture and all the reagents are maintained on ice throughout the preparation and dispensing of the mixture into the tubes and the addition of the template to the mixture. While this is done, a soak cycle of 80°C for five minutes is programmed into the cycler. When the heat block of the thermocycler reaches 80°C, the tubes are removed from the ice and placed immediately on the heat block. From then on, the thermal cycling protocol proceeds normally.

**Sequencing of amplification products**

There are many suitable protocols available for determining nucleotide sequencing of PCR products, both from clones or directly from PCR reactions. These include single and double-stranded template sequencing with Sequenase (USB, Cleveland, OH), cycle sequencing, and other techniques utilizing thermostable DNA polymerases. Each has its advantages which must be evaluated by the investigator as best suited for the intended goal of the project.

It should be noted, however, that direct sequencing of PCR products normally yields a 'consensus' sequence as the PCR product represents a 'pool' of individual amplicons reflecting both template variation, template integrity, and polymerase fidelity. Sequences proceeding from cloned amplicons represent the sequence of that single amplicon ligated to the vector. It is not a consensus sequence and might reflect both template variations and/or polymerase errors. When sequencing cloned amplicons it is recommended that a minimum of six different clones be used to generate a 'consensus' sequence. Alternatively, purified plasmid DNA from 10-20 clones may be pooled and a single sequencing reaction conducted as this represents a 'consensus' sequence of the 10-20 clones pooled.

Automation of DNA sequencing including the incorporation of fluorescent dye chemistry has greatly improved both the quality and output capabilities of sequencing DNA from all sources. In dye terminator chemistry, fluorescent tags are attached to the chain terminating nucleotides with each of the four dideoxynucleotides carrying a spectrally different fluorophore. During cycle sequencing both dye labelled dideoxynucleotides and deoxynucleotides are present, resulting in random chain termination during nucleotide incorporation and labelled molecules of almost every possible base length. Unincorporated dye-terminators are then removed from the reaction by using a spin column or ethanol precipitation step. Each reaction is subsequently electro-phoresed on a polyacrylamide sequencing gel, utilizing only one lane on the gel for each primed reaction. The labelled DNA fragments are detected by their fluorescence as they migrate past the detector which scans horizontally across the gel.

Automated fluorescent DNA sequencing systems in general offer many advantages over manual sequencing in accuracy, reproducibility, and ease of use. Both the software and basic chemistry used in automated fluorescence sequencing have drastically improved sequence quality and output. The most significant advantage of this system is the ability of computer software to perform base-calling and sequence analysis, eliminating the possibility of errors arising when DNA sequences are read and processed manually. Automation also permits one to easily and qualitatively compare multiple ratios of the same sequence for determining consensus sequences or positioning. Analysis software allows the review of the run conditions (voltage, wattage, amberrage and temperature) and error and command logs providing validation and trouble-shooting of each run. Furthermore, sequence assembly software can 'clean up' sequences, identifying and removing ambiguous stretches and primer or plasmid sequences.

Several modifications of fluorescent cycle sequencing chemistry have assisted in improving automated DNA sequencing and base-calling. Improvements specifically in dye terminator chemistry include the incorporation of dITP in place of dGTP in the reaction mix (Perkin Elmer/Applied Biosystems Division, Foster City, CA) which aids in minimizing band compressions for more accurate base-calling. Five per cent DMSO included into sequencing mixes increases peak height variability, contributing to base-calling precision. Development of Taq polymerase FS (Perkin Elmer/Applied Biosystems Division, Foster City, CA), a variant of Taq DNA polymerase, has further refined fluorescent dye cycle sequencing processes. Mutations in the active site and the amino terminal domain result in less discrimination against the labelled terminating dideoxynucleotides and decreased 5'-3' exonuclease activity, respectively. The end result of these mutations is a much more even peak intensity pattern leading to more accurate base-calling by the software.

Aside from the general advantages described above, automated fluorescent sequencing has several benefits when specifically working with ancient DNA. Accurate sequence identification is especially crucial when sequencing ancient DNA, for both classification purposes and phylogenetic analysis. Manual base-calling cannot always be completely eliminated with automated fluorescent systems, but the degree to which base-calling relies on perception is greatly diminished. Assignment of International Union of Biochemistry (IUB) codes and viewing electropherograms of aligned forward and reverse strands can also effectively identify potential polymorphisms important in phylogenetic studies. Heterozygous base positions are more easily observed by viewing an electropherogram created by automated fluorescent system software. Computer software programs such as Factura (PE/ABI) assign IUB codes to mixed base positions by using a ratio set by the user to compare the highest peak with each of the other three peaks in the same location. If the ratio between any of the two lower peaks and the higher one is above the set threshold percentage, an IUB code is assigned. This aids in the detection of heterozygous positions (Figure 5), which may be present at certain positions in multi-copy genes (for
Figure 5 Factura™ output from a dye dideoxy terminator cycle sequencing reaction indicating the presence of a heterozygous position (see arrow) in the 16S rRNA of ancient Bacillus spp.

example, 16S rRNA gene) and undetectable by autoradiographic methods.

Regardless of the method used and the approach to sequencing, sequence reproducibility and comparison with those of known, related taxa should be performed to increase the degree of reliability on the sequence data generated from ancient DNA.

References