Taxonomy, DNA, and the Bar Code of Life

MARK STOECKLE

DNA sequence analysis is enormously useful in studies of evolutionary history. Extensive sampling of DNA sequences has helped establish the diversity of life and allowed researchers to analyze evolutionary relationships within groups in detail. DNA sequencing has also been applied to identify specimens and resolve species boundaries in populations of apparently similar organisms. However, the bewildering variety of genes and methods of analysis employed in DNA-based phylogenetic and identification research has generally limited the applicability of results beyond the specific groups under study. The potential utility of a large-scale effort to sequence uniform gene targets across all species of life was the subject of "Taxonomy and DNA," a conference held at Cold Spring Harbor Laboratory on 9–12 March 2003. The conference participants included specialists in animal, microbial, and plant taxonomy; molecular biology; and bioformatics. The goals of a large-scale sequencing project are to enable a practical method for species identification and to provide insight into the evolutionary history of life.

Is species identification through DNA analysis possible? From a theoretical point of view, a remarkably short DNA sequence should contain more than enough information to resolve 10 million or even 100 million species. For example, a 600-nucleotide segment of a protein-coding gene contains 200 nucleotides that are in the third position within a codon. At these sites, substitutions are usually selectively neutral, and mutations accumulate through random drift. Even if a group of organisms had only either adenosine or thymine (or, alternatively, only guanosine or cytosine) at third nucleotide positions—an unrealistically pessimistic assumption—there would still be 2^{200}, or 10^{60}, possible sequences of third-position nucleotides. In reality the number is likely to be far higher.

The main limiting factor in distinguishing closely related species is likely to be the rate of accumulation of new mutations. DNA sequence analysis of a uniform target gene to enable species identification has been termed *DNA bar coding,* by analogy with the Uniform Product Code bar codes on manufactured goods. Proof that DNA bar coding can distinguish at least some species has been provided by analysis of cytochrome c oxidase subunit I (COI) sequences among closely related species across diverse phyla in the animal kingdom (Hebert et al. 2003).

What are the benefits of DNA bar coding? As a uniform, practical method for species identification, it appears to have broad scientific applications. It will be of great utility in conservation biology, for example, including biodiversity surveys. It could also be applied where traditional methods are unrevealing: identification of eggs and larval forms, for instance, and analysis of stomach contents or excreta to determine food webs. DNA-based species identification will help open the treasury of biological knowledge, which is currently underused partly because taxonomic expertise for species identification is relatively inaccessible. DNA bar coding can be likened to aerial photography, in that it provides an efficient method for mapping the extent of species, though in sample space rather than physical space. The "aerial map" of DNA bar codes will help investigators explore the biological world and make full use of the enormous knowledge that has been built on 250 years of classical taxonomy. And as sequencing costs decrease, DNA-based species identification will become available to an increasingly wide community. When costs are low enough, science teachers and backyard naturalists will be able to use DNA bar coding for in-depth examination of local ecosystems.

The number of living species on Earth is unknown. There are approximately 1.7 million named species and possibly another 10 million (not counting bacteria and archaea) that have not been described. DNA-based identification could be vitally important in flagging specimens that represent undescribed taxa. Comprehensive analysis of populations will help identify cryptic species, which may be far more prevalent than commonly realized, even among large animals. DNA sequencing of the same gene (or set of genes) across diverse phyla will help unravel the processes that underlie speciation and reveal the diversity of life to an extent that is not now possible.

What are the limitations of DNA bar coding? DNA-based species identification depends on distinguishing intraspecific from interspecific genetic variation. The ranges of these types of variation are unknown and may differ between groups. It may be difficult to resolve recently diverged species or new species that have arisen through hybridization. There is no universal DNA bar code gene, no single gene that is conserved in all domains of life and exhibits enough sequence divergence for species discrimination. The validity of DNA bar coding therefore depends on establishing reference sequences from

taxonomically confirmed specimens. This is likely to be a complex process that will involve cooperation among a diverse group of scientists and institutions.

What genes are appropriate targets for DNA bar coding? The ideal gene target is sufficiently conserved to be amplified with broad-range primers, yet divergent enough to resolve closely related species. For animals, mitochondrial genes are attractive targets because they are shared across diverse taxa and do not contain introns that can complicate amplification using the polymerase chain reaction (PCR). The availability of broad-range primers for amplification of mitochondrial COI from diverse invertebrate phyla establishes this gene as a particularly promising target for species identification in animals (Folmer et al. 1994).

Different gene targets or protocols may be needed for certain taxonomic groups, but there is no apparent barrier to applying some type of bar coding across all domains of life. Plants have relatively little sequence variation in their mitochondrial DNA, perhaps because of hybridization and introgression. A chloroplast gene such as *matK* (maturase K) or a nuclear gene such as *ITS* (internal transcribed spacer) may be an effective target for bar coding in plants. The mitochondrial genes of fungi contain many introns, but this limitation can be circumvented by employing reverse transcription in conjunction with PCR. The feasibility of DNA bar coding of protists via COI has not been examined in depth. In the case of bacteria and archaea, the small subunit ribosomal RNA gene (SSU rRNA) can be used for determination of major groups, and additional genes will be needed for more detailed resolution. Indeed, a DNA bar coding project could with advantage include SSU rRNA sequencing of all specimens, as comparison of SSU rRNA sequences is the basis for the existing Tree of Life project and other investigations into deep evolutionary relationships.

What would a Bar Code of Life project cost? The cost of a comprehensive sequencing project depends largely on the cost of obtaining DNA from taxonomically verified specimens. It may be possible to use museum collections, including specimens preserved in formalin (Fang et al. 2002). Institutions with comprehensive collections of one or more species groups could be the launching pads. Amplification and sequencing are inexpensive, costing about $1 per specimen. The resultant flood of sequences could be made publicly available on the Internet. There may be a need for a stand-alone, curated database to supplement GenBank; the stand-alone database would be designed to integrate sequence data with specimen and taxonomic information. Once sequence divergences are characterized, other technologies, such as chip-based DNA arrays, could be applied for species recognition and might be usable in field biology. A large-scale sequencing project is likely to drive development of faster, better, and cheaper technologies for DNA sample preparation and analysis, just as the Human Genome Project enabled the development of robotic sequencing.

Establishing a Bar Code of Life program will involve the cooperation of a diverse group of scientists and institutions. Although bar coding is not optimal for the study of deep evolutionary relationships, DNA-based species identification offers enormous potential benefits for the biological scientific community, educators, and the interested public. It will help open the treasury of biological knowledge and increase community interest in conservation biology and understanding of evolution.

Mark Stoeckle (e-mail: MarkStoeckle@nyc.rr.com) is Guest Investigator in the Program for the Human Environment, Rockefeller University, New York 10021.

References cited

Fang SG, Wan QH, Fijihara N. 2002. Formalin removal from archival tissue by critical point drying. BioTechniques 33: 604–611.

Folmer O, Black M, Hoeh W, Lutz R, Vrijenhoek R. 1994. DNA primers for amplification of mitochondrial cytochrome c oxidase subunit I from diverse metazoan invertebrates. Molecular Marine Biology and Biotechnology 3: 294–299.

Hebert PDN, Cywinska A, Ball SL, deWaard JR. 2003. Biological identifications through DNA barcodes. Proceedings of the Royal Society of London, B 270: 313–322.