First publ. in: Trends in Ecology and Evolution 9 (1994), 8, pp. 278-280

Shortcomings of the cytochrome *b* gene as a molecular marker

The widespread successful application of the polymerase chain reaction (PCR) to questions in animal ecology and evolutionary biology can be attributed partly to the discovery of so-called 'universal', or better, 'versatile' PCR primers (e.g. Refs 1,2) for mitochondrial DNA (mtDNA). The mtDNA molecule is by far the most commonly used in population and evolutionary biology since it features several advantages (such as higher rates of evolution) that make it the usual choice for population-level questions[3,4]. Despite the generally much-elevated rates of evolution of mtDNA compared to nuclear DNA, conserved regions can also be found that are invariant across a wide taxonomic range. The mitochondrial versatile primers are located in these regions and their conservative unspecific DNA sequences permit the determination of DNA sequences for many groups of animals for which no DNA sequence information previously existed. Versatile PCR primers allowed immediate access to the mitochondrial genome for the novice molecular systematist since no highly technical pilot work is necessary. The use of the same set of primers by many laboratories created a universal metric, which is easily stored in databanks since homologous DNA sequences permit comparisons among studies and species.

Versatility of cytochrome *b*

Five years ago, the first set of versatile primers was published for portions of the cytochrome *b* gene, the small ribosomal RNA gene, and the major noncoding region. These primers were found to amplify homologous portions of mtDNA in all classes of vertebrates and many other animal phyla[1]. The primers for the cytochrome *b* gene gained immediate and widespread acceptance and have been used in a large number of studies, with great success, for a wide range of questions in ecology and evolution (e.g. reviewed in Refs 4,6), particularly for studies on many vertebrates (in invertebrates and primates the mitochondrially encoded cytochrome oxidase subunits are used most often, e.g. Ref. 5). Cytochrome *b* is the most widely used gene for phylogenetic work. It has become somewhat of an 'industry standard' since the first publications seemed to suggest that it would provide answers to every question: it was thought to be variable enough for population questions (e.g. Ref. 7) and conserved enough for 'deep' phylogenetic questions among distantly related organisms (mammals[8], birds[9,10], classes of vertebrates[1,11]).

There are several good reasons for the continued use of cytochrome *b* as a phylogenetic marker. Although it is slow in terms of amino acid substitutions, the rate of evolution for silent substitutions in third codon positions is similar to that of other mitochondrial genes. Because of its widespread use and its status as a universal metric, results of particular studies can be meaningfully compared with a larger body of work. It is probably the best-known mitochondrial gene with respect to its structure and function (e.g. Ref. 6). This point is important, since increasing sophistication in phylogenetic analyses of DNA sequences will also lead to more-detailed consideration of functional constraints on the gene product. Typically, variation in amino acid residues is found in membrane-spanning domains of this molecule. However, several recent publications point out that the expectations in cytochrome *b* as a molecular marker in ecology and evolution may have been too high and that this gene may not hold all the answers after all[7,12–16].

Problems with cytochrome *b*

No gene can be expected to be perfect for all questions, and cytochrome *b* is no exception. The problems that have been encountered when using cytochrome *b* include base compositional biases, rate variation between different lineages, early saturation of third codon positions, and limited variation in first and second codon positions resulting in little phylogenetic information for 'deep' evolutionary questions. Moreover, the recent discovery of the presence of some nuclear copies of this gene in some birds and mammals has made this a cause for concern for phylogeny reconstruction[10,17,18]. Some of cytochrome *b*'s shortcomings can be circumvented by using other

genes; others are general properties of mtDNA and no easy way out can be offered.

It might be obvious, but at the planning stage of a study one needs to keep an open mind, consider several alternative markers, and be informed about the ways of dealing with potentially expensive and time-consuming problems. Before plunging into a large-scale sequencing project, it behoves one to collect preliminary data. Species represented in this preliminary study should include the ones expected to be most distantly related and the ones expected to be most closely related, to test for the appropriateness and the level and mode of variation of the considered gene. Preliminary tests should include bootstrap tests for the robustness of the phylogenetic results, tests of levels of homoplasy of characters, transition–transversion ratios to test for saturation of the gene, relative rates tests, etc.

Cytochrome *b* may be a good start because reliable primers are around, but it may not be the best choice in the end. The mitochondrial genome of animals typically contains 13 protein coding genes, two genes coding for ribosomal RNA, 22 transfer RNA (tRNA) genes and one large non-coding region. Versatile PCR primers are available for most of these genes and hence exploration is feasible[2,16]. Other mitochondrial candidate genes offer some of the same advantages as cytochrome *b*, or might be better choices for particular projects. However, some of the potential shortcomings of cytochrome *b* also hold for other conservative mitochondrial genes (e.g. cytochrome oxidases).

Genes that show more variation in terms of amino acid changes between species such as ATPase genes will, by their very nature, make it more difficult to design primers with a high degree of versatility. Nonetheless, tRNA genes, which tend to be quite conservative, are interspersed between protein-coding genes, and will allow the design of versatile primers with which one can gain access to the more variable mtDNA regions. However, the ribosomal genes and even the more-quickly evolving protein-coding genes, such as ATPase 6 and 8, and the major non-coding region are not going to be ready substitutes for cytochrome *b* in all studies. Because the first set of versatile primers published was for the 5′ end of the cytochrome *b* gene[1], the other end has been determined less frequently. This part of the gene is generally more variable in terms of amino acid variation and might provide more useful phylogenetic information for questions among distantly related species[6]. Generally, protein-coding genes have the important advantage over the ribosomal genes that alignment is not problematic.

Transversions and substitutions that result in amino acid changes accumulate more linearly than transitions in third positions, and are hence more reliable markers for phylogenetic questions among distantly related organisms. Transitions occur up to 20 times more often than transversions and back-mutate frequently, which makes them unsuited for 'deep' phylogenetic questions. The conserved nature of the cytochrome *b* gene at the amino acid level will result in reduced levels of variation in first and second codon positions, where most nucleotide substitutions will cause amino acid changes. This makes these changes more reliable markers for distant relationships because back mutations are less likely, but it also lowers the amount of phylogenetic information available among distantly related species[9,11–13]. Some conservative amino acid substitutions (e.g. leucine–isoleucine, leucine–valine) will occur rapidly and back-mutate, and hence may not be reliable for phylogenetic inference among distantly related species and should be excluded from the analyses[8]. Transitions in third codon positions should be down-weighted or excluded from the phylogenetic analysis among distantly related species since they can potentially be phylogenetically disinformative[8,9,11,15]. Also, leucine codons can have silent mutations in first codon positions (leucine is coded for by two codon families, UUA/G and CUN, where N represents any base) and these first positions should be treated equal to silent third position mutations, that is, down-weighted in parsimony analyses or ignored by using transversions only[8,15].

Rate variation in cytochrome *b* in terms of amino acid substitutions was observed early[1] and is now believed to be a general property of the whole mitochondrial genome[19,20]. Endotherms usually have a faster rate than ectotherms[19,20]. Variation in the rate of molecular evolution has the potential to obscure phylogenetic signals and to hinder the recovery of the true evolutionary relationships. A recent study comparing all mitochondrial protein-coding genes of the opossum with those of placental mammals (their split occurred about 80 million years ago) described particularly unclock-like behavior for cytochrome *b* (and the cytochrome oxidase genes)[15]. The NADH genes were found to evolve at much more even rates among mammals, and might hence be more-reliable phylogenetic markers than cytochrome *b* (Ref. 15). Cytochrome *b* also seems to evolve at different rates in different lineages of amphibians: the frogs' rate appears to be slower than that of salamanders[13]. These recent findings enforce the notion that even within endotherms and ectotherms noticeable rate variation has to be expected and might be enough reason to avoid cytochrome *b* as a phylogenetic marker among distantly related species[13,15]. Rate-robust phylogenetic methods might provide the only remedy on how to deal with this apparent shortcoming of cytochrome *b*.

Base compositional biases

Base compositional biases are particularly pronounced at third codon positions where substitutions tend to be silent – and they are not particular to the cytochrome *b* gene. Third positions are most likely to be variable among populations and closely related species. Base compositional biases will lower the detected amount of variation and might lead to disappointments with cytochrome *b* for population work, since it might contain little scored variation. This has led to the initially surprising observation that third positions, because of base compositional biases and slow rates of amino acid changes, might saturate rather quickly and that variation accumulates slowly in the other positions. All mitochondrial genes that are encoded on the H-strand are low in guanine, but the other three bases are often not found in equal proportions either; cytosine tends to be the most frequent nucleotide (e.g. in birds)[1]. In third codon positions, quite commonly a bias is found not only against guanine but also against thymine. In larger taxonomic groups, like fishes, both general types of biases are found in third positions (anti-guanine and anti-guanine + anti-thymine) but these do not seem to occur in an obvious phylogenetic pattern[16]. When one addresses phylogenetic questions at levels where third codon positions are saturated with back mutations (typically 15–20% overall uncorrected sequence divergence) and do not provide reliable phylogenetic information, not enough variation is found in first and second codon positions either. This is where cytochrome *b* is at its weakest and most frustrating and where other, more-variable protein-coding genes (e.g. NADH or ATPase genes) should be used.

Insects generally have highly A+T-rich mitochondrial genomes (for example, the cytochrome *b* gene of the honeybee, *Apis mellifera*, is more than 80% A+T[21]). Extreme base compositional biases are responsible for strong codon biases; several codons are not used at all[21]. These biases will tend to increase differences between species in terms of amino acid differences because the only 'allowed' bases at third codon positions will involve transversions that sometimes

result in residue changes. Also, amino acid substitutions that are caused by transversions in second and first codon positions will tend to be less conservative than residue substitutions that are caused by transitions. Base compositional biases are potentially harmful to phylogeny reconstruction: they will lead to less variation at the DNA level because the number of character states is reduced, increasing the occurrence of homoplasy in the data set (causing a low ceiling for saturation). These effects will be particularly difficult to deal with if the species under consideration differ in the direction of their biases and independently converged on the same bias[22].

Potential of the major non-coding region

Some portions of the major non-coding region have functional importance and change slowly. For population-level work, several primers, which have various levels of 'versatility' and are located in these conserved regions, have been published (e.g. Refs 1,23,24). This mitochondrial region generally accumulates mutations at rates 3–5 times the average rate in cytochrome *b*, and provides much higher amounts of information at the population level, where cytochrome *b* was found to be too slow (e.g. Refs 7,25). Sequencing the major non-coding region on average might be comparable to sequencing a protein-coding gene that is entirely composed of third positions. Therefore, the advantage of this region for population-level questions is that it will provide much detailed genealogical information. Interestingly, the major non-coding region actually seems to be reliable up to quite some phylogenetic depth[23,24]. However, the rate of evolution of the major non-coding region, just as for cytochrome *b*, seems to vary between different groups of organisms (e.g. salmonid fishes seem to be slower than expected and bill fishes seem to be faster, based on cytochrome *b* variation; Meyer *et al.*, unpublished). One disadvantage that the major non-coding region has, relative to cytochrome *b*, is that alignment and therefore positional homology across studies might not be easily established. Nonetheless, although it may not be a universal metric, it would seem justifiable to make comparisons across studies if, for example, the region between the proline-tRNA and the D box is compared. Despite these drawbacks, the non-coding region still has advantages over cytochrome *b* for population-level questions.

The abundant use of DNA sequence data in ecology and evolutionary biology is an exciting new development. For laboratories that set out to undertake molecular phylogenetic work and need to produce preliminary data for grant proposals, the cytochrome *b* gene is often the first choice. However, the appeal of cytochrome *b* as an easy 'beginner's gene' for phylogenetic work is tarnished by several of its particular shortcomings. One should not expect that this gene is going to be the right gene for all questions. More appropriate alternatives exist and should be seriously considered.

Acknowledgements

I thank S. Edwards, G. Ortí and P. Ritchie for discussion and the National Science Foundation for support.

Axel Meyer

Dept of Ecology and Evolution, and Program in Genetics, State University of New York, Stony Brook, NY 11790-5245, USA

References

1 Kocher, T.D. *et al.* (1989) *Proc. Natl Acad. Sci. USA* 86, 6196–6200
2 Palumbi, S.R. *et al.* (1991) *Simple Fool's Guide to PCR*, University of Hawaii Press
3 Wilson, A.C. *et al.* (1985) *Biol. J. Linn. Soc.* 26, 375–400
4 Avise, J.C. (1994) *Molecular Markers, Natural History and Evolution*, Chapman & Hall
5 Adkins, R.M. and Honeycutt, R.L. (1994) *J. Mol. Evol.* 38, 215–231
6 Esposti Degli, M. *et al.* (1993) *Biochem. Biophys. Acta (Bioenergetics)* 1143, 243–271
7 Wenink, P.W., Baker, A.J. and Tilianus, M.G.J. (1993) *Proc. Natl Acad. Sci. USA* 90, 94–98
8 Irwin, D.M., Kocher, T.D. and Wilson, A.C. (1991) *J. Mol. Evol.* 32, 128–144
9 Edwards, S.V., Arctander, P. and Wilson, A.C. (1991) *Proc. R. Soc. London Ser. B* 243, 99–107
10 Kornegay, J.R. *et al.* (1993) *J. Mol. Evol.* 37, 367–379
11 Meyer, A. and Wilson, A.C. (1990) *J. Mol. Evol.* 31, 359–364
12 Normark, B.B., McCune, A.R. and Harrison, R.G. (1991) *Mol. Biol. Evol.* 8, 819–834
13 Graybeal, A. (1993) *Mol. Phyl. Evol.* 2, 256–269
14 Hillis, D.M. and Huelsenbeck, J.P. (1992) *J. Hered.* 83, 189–195
15 Janke, A. *et al.* (1994) *Genetics* 137, 243–256
16 Meyer, A. (1993) in *Molecular Biology Frontiers, Biochemistry and Molecular Biology of Fishes* (Vol. 2) (Hochachka, P.W. and Mommsen, T.P., eds), pp. 1–38, Elsevier
17 Smith, M.F., Thomas, W.K. and Patton, J.L. (1992) *Mol. Biol. Evol.* 9, 204–215
18 Quinn, T.W. (1992) *Mol. Ecol.* 1, 105–117
19 Martin, A.P. and Palumbi, S.R. (1993) *Proc. Natl Acad. Sci. USA* 90, 4087–4091
20 Rand, D.M. (1994) *Trends Ecol. Evol.* 9, 125–131
21 Jermiin, L.S. and Crozier, R.H. (1994) *J. Mol. Evol.* 38, 282–294
22 Steel, M.A., Lockhardt, P.J. and Penny, D. (1993) *Nature* 364, 440–442
23 Meyer, A. *et al.* (1990) *Nature* 347, 550–553
24 Meyer, A., Morrissey, J.M. and Schartl, M. (1994) *Nature* 368, 539–541
25 Edwards, S.V. and Wilson, A.C. (1990) *Genetics* 126, 695–711