A plea for DNA taxonomy

Diethard Tautz[1], Peter Arctander[2], Alessandro Minelli[3], Richard H. Thomas[4] and Alfried P. Vogler[5]

[1]Institut für Genetik der Universität zu Köln, Weyertal 121, 50931 Köln, Germany
[2]Department of Evolutionary Biology, Institute of Zoology, University of Copenhagen, Universitetsparken 15, DK-2100 Ø Copenhagen, Denmark
[3]Department of Biology, University of Padova, Via Ugo Bassi 58 B, 35131 Padova, Italy
[4]Department of Zoology, The Natural History Museum, Cromwell Road, London SW7 5BD, UK
[5]Department of Entomology, The Natural History Museum, Cromwell Road, London SW7 5BD, UK

Taxonomy underpins all biological research, with implications for many basic scientific and applied fields. Insights into the stability or change of animal and plant guilds require species identification on a broad scale and biodiversity questions have become a major public issue. But this comes at a time when taxonomy is facing a crisis, because ever fewer specialists are available. Here, we explore the possibility of using DNA-based methodology to overcome these problems. The utility of DNA sequences for taxonomic purposes is well established. However, all current taxonomic approaches intend to use DNA, at best, as an auxiliary criterion for identifying a species or a taxon, but have not given it a central role. We propose a scheme in which DNA would be the scaffold of a taxonomic reference system, whilst maintaining the importance of the morphological information associated with whole specimens.

Current taxonomy represents a body of work that has accumulated over the past ~250 years, since the introduction of the binomial naming system by Linnaeus in the 1750s [1,2]. Specific rules have been established for recognizing, naming and classifying species* to avoid redundant descriptions or the use of the same name for more than one species. These rules were introduced in the late 19th century and are continuously monitored by international commissions of scientists (http://www.iczn.org and http://www.botanik.univie.ac.at/iapt/). A crucial component of current practice in taxonomy is the concept of the type specimen that serves as the central reference for comparisons. Designating a type specimen is required when a new species is named, and these are usually deposited in major museum collections, where they are available for study. This system, with its main anchor in publicly funded collections, has produced a reliable and steadily updated taxonomy.

However, the system depends heavily on specialists whose knowledge is frequently lost when they retire. Furthermore, there is a clear bias of focus on particular groups, such as vertebrates, insects or flowering plants, whilst other important groups, such as nematodes, mites or diatoms, are neglected. Finally, although most of the specialist knowledge is published in some form, the respective literature is often difficult to access. Web-based technology could be a great step towards a more accessible and universal platform for the deposition and retrieval of taxonomic information. We endorse such moves in principle, as they will greatly speed up communication, and make species diagnoses and new descriptions more accessible [3,4]. However, several problems remain, not least the quality and accuracy of the submitted information [5].

Here, we propose to introduce a DNA taxonomy system to provide a new scaffold for the accumulated taxonomic knowledge and as a convenient tool for species identification and description [6]. Although the following discussion will focus primarily on the advantages of a DNA-based taxonomy system, it is not meant to be a critique of morphology-based taxonomy. We are very much aware of the strengths of the current system and we believe therefore that a DNA-based system must be firmly anchored within the knowledge, concepts, techniques and infrastructure of traditional taxonomy.

DNA-based taxonomy

The basic procedures of DNA taxonomy would be straightforward. A tissue sample is taken from a collected individual and DNA is extracted from this. This DNA serves as the reference sample from which one or several gene regions are amplified by PCR and sequenced. The resulting sequences are, as a first approximation, an identification tag for the species from which the respective individual was derived. This sequence is made available via appropriate data bases, together with the species description and other associated information, ideally including its taxonomic status with appropriate references. The sequence now serves as a standard for future reference, together with the type specimen and the respective DNA preparation, which will be deposited in museum collections. Once a significant sequence data base has been built up, new samples can be checked against these existing sequences to assist species re-identification or to assess whether a new species description might be warranted. The data base could also serve to resolve questions about the taxonomic identity of specimens that are derived from larval life stages, or for identification of artefacts from trade with endangered species and so on. In

Corresponding author: Diethard Tautz (tautz@uni-koeln.de).
* Here, we do not refer to bacterial taxonomy, because bacteria have to be treated differently in many respects [16] and because schemes for DNA taxonomy in bacteria are already being developed (http://www.dsmz.de/bactnom/bactname.htm).

 0169-5347/02/$ - see front matter PII: S0169-5347(02)00041-1

Box 1. Limits of DNA taxonomy

It must be emphasized that the power of DNA sequences for identifying species is limited when species pairs have very recent origins. For some time after the initial split, new sister species will share alleles, either because of ongoing gene flow, or because of recent ancestry. In such cases, sequences from one or few individuals will not be sufficient for an unequivocal assignment to a particular group. There is also a special complication for organelles (mitochondria or chloroplasts), which can occasionally be transferred, at least between closely related species. This could result in different diagnoses, depending on whether one uses a sequence from the nuclear genome or from the organelle genome.

The buildup of sequence differences that can serve as unequivocal characters depends on the mutation and fixation rates. The combined rates for neutrally evolving sites are between 0.1% and 2% per million years in nuclear sequences and can reach up to 5% in mitochondria. The random fixation for a new mutation is expected to occur within $4N_e$ generations for nuclear loci and within $1N_e$ generations for mitochondria (N_e is the effective population size and measures only the number of reproductively effective individuals, which is usually much lower than the census size). This provides some guideline for assessing how long it will take until one can expect to find a diagnostic difference between newly evolved species after cessation of gene flow. Although this time should be assessed on a case-by-case basis, it will usually not be more than 100 000 years. It seems safe to assume that most currently described species are older than this.

Evidently, there are special cases, such as the very recent radiation of cichlid fishes in Lake Victoria, where the morphological distinctiveness has built up much faster than has the molecular one. Morphology-based taxonomy is clearly more powerful in such cases. Still, DNA analysis is not useless, because it provides essential insights into the time frame of the radiation and the origin of the colonizing animals [a]. Moreover, analysis of allele frequency changes, rather than diagnostic changes, enables us to identify very recent divergence of species, even under sympatric conditions [b].

References

a Meyer, A. *et al.* (1990) Monophyletic origin of Lake Victoria cichlid fishes suggested by mitochondrial DNA sequences. *Nature* 347, 550–553

b Schliewen, U. *et al.* (2001) Genetic and ecological divergence of a monophyletic cichlid species pair under fully sympatric conditions in Lake Ejagham, Cameroon. *Mol. Ecol.* 10, 1471–1488

the early phases of this initiative, concerted efforts must be made to achieve good coverage of all known species (or some specifically targeted subgroups), but once the data base is sufficiently complete, these comparisons, aided by the phylogenetic analysis of query sequences, will readily place any sequences from new specimens.

DNA sequence information is digital and is not influenced by subjective assessments. It would be reproducible at any time and by any person, speaking any language. Hence, it would be a universal communication tool and resource for taxonomy, which can be linked to any kind of biological or biodiversity information. Even if a query sequence does not produce an exact match, it will be possible to link an organism to closely related ones. Although DNA taxonomy has limitations (Box 1), it would have the advantage of being a universally applicable tool.

Naming of species

A Linnaean name is an anchor for biological information about a species, including its taxonomic affinities, morphology, distribution and possible ecological role. Biologists need to use species names for communicating with each other. Nucleotide strings cannot serve these purposes and a DNA-based system will therefore require a separate naming system.

The binomial Linnaean naming system (comprising the genus and species name) is well established and broadly used, but it is inherently unstable. It requires that a species is associated with a particular genus, but this association is only a hypothesis, which can change when new data become available. Thus, a species originally named *Arbitrarus conventicus* could become *Revisionus conventicus*, when new species are identified, or new data are evaluated that justify its inclusion into the genus *Revisionus*. In fact, much of the taxonomic literature includes discussion of such cases. A name that has been used for a long time thus can suddenly disappear and only specialists might eventually be able to identify its fate.

If DNA sequences serve as the main reference, name changes become less of a problem, because the sequence will always provide the link to the previously used name. Thus, the nomenclatural instability in the established Linnaean naming system would be ameliorated and the formal rules governing the naming under the Linnaean system can be retained. Only the convention to refer to the author who has first described the species should be extended to include the reference to a numbering system that refers to the respective data base entry. This would then be akin to the situation for humans, where names are used for communication, but passport numbers or social security numbers are used for unequivocal identification [7].

Matching Linnaean names with DNA sequences

We propose that an attempt is made to provide a DNA sequence alongside all future taxonomic samples and species descriptions, a need that is well recognized in contemporary studies [8]. This should not be a technical problem, particularly when appropriate facilities are established (see below). However, the real challenge for DNA-based taxonomy is to provide a particular DNA sequence for the species that have already been named. Ideally, the DNA information would be obtained from the type specimen itself, but this will be impossible in most cases, either because types are not available or because they cannot be used for DNA extraction. In such cases, DNA taxonomy will have to be based mainly on sequences from newly collected individuals, which are assessed by experienced taxonomists to determine their identity. This specimen and associated sequence then provides a reference record and all further sequences that are very similar or identical could be associated with the same name or data base entry. In selecting the respective specimen, we would recommend following the policy as suggested in the International Code of Zoological Nomenclature [9], when the original type specimen has been lost, or is otherwise unavailable. In these instances, taxonomists can fix the species name by selecting a replacement

type, or neotype. The advice of the International Code is to select specimens from the original type locality and that are in good agreement with the original description of the species.

The role of collections

A biologist wants to see a preserved organism as a whole and to retain as much of it as possible. However, the integrity of specimens cannot always be guaranteed, even in morphological studies. We would argue that one should be prepared to accept damage or destruction of specimens for DNA taxonomy. For large animals and for most plants and fungi, this is usually not a problem, because only a small portion of a specimen will suffice. DNA extractions from insects can also be performed nondestructively. But there will also be specimens that have to be fully destroyed to extract sufficient DNA. In these cases, possibly the only way to preserve at least some morphological information would be to photograph the specimen before destruction.

The essential reference object in a DNA taxonomy scheme would be the DNA sample that is obtained from the type specimen. Collection and curation of extracted DNA samples is technically easy. DNA is very stable either in a buffered solution, as ethanol precipitate, or freeze dried. As DNA samples of any organism have the same storage requirements, one could shelve them simply in the order in which they have been obtained, simplifying the organization of collections. Any sample should be split into multiple subsamples, which could then be distributed among various museums, or sent to researchers for further study.

Some samples might eventually gain commercial value, as they might include genes of economic importance for biomedicine or agriculture. It will therefore be necessary to establish a tracking system that enables us to assign legitimate ownership of the samples (according to the Convention on Biological Diversity, this could be the country of origin).

The need for establishing standards for DNA collections is already growing rapidly. Although there are many projects using DNA analysis for studying the phylogeny or phylogeography of species, there is currently no unified scheme for voucher deposition (i.e. sampling species). Safeguarding these potentially valuable samples for future reference should be one of the most urgent goals.

Sequences to be used

Although any part of the genome of an organism provides us with some information about its taxonomic affiliation, some regions are more useful than others. The genes with the broadest taxonomic coverage currently available are those encoding the ribosomal small subunit sequences, both of nuclear and mitochondrial origin. However, this is a rather conservative gene, which is not particularly useful for differentiating closely related species. One of the most quickly diverging, and thus very informative sequences, is the mitochondrial control region but this is not very useful for determining higher taxonomic affiliations. Mitochondrial *Cytochrome b* gene sequences have also been used extensively, particularly for vertebrates [10]. They enable researchers to resolve relationships between closely related taxa as well as to construct higher level phylogenies, because the synonymous third codon positions evolve fast, whereas the protein sequence, as such, evolves relatively slow. This approach has been successfully explored by Hebert *et al.* [11] (see also J. Mallet, this issue).

However, mitochondrial transfer can occur between closely related taxa and copies of mitochondrial genes frequently have been transposed to the nucleus, potentially creating confusion of provenance. A reasonable alternative could be the divergence loops in ribosomal DNA (rDNA) sequences, in particular those of the large subunit rRNA. These are faster evolving, but embedded in regions that are more conserved, which enable the use of almost universal primers. Moreover, because ribosomal genes are pre-amplified in all organisms, they should also be more easily retrievable from very small or partially degraded samples. Finally, rRNAs are so abundant in cells that they could serve directly as probes for DNA-microarray approaches for species identification [12].

In any case, it seems advisable to use more than one sequence region for assigning taxonomic status. This might also give hints for possible hybrids that would need to be further analyzed. It seems probable that some taxon-specific preferences will be developed and will be followed by the specialists working with the respective group [13]. A universal agreement about the type of molecules and genes to be analyzed does not seem to be necessary.

Taxonomy and phylogeny

The purpose of taxonomy is the identification of species and their assignment to higher level taxa. The latter is often associated with generating phylogenetic hypotheses, which can potentially be inferred directly from DNA sequences. Although the sequences collected within the framework of DNA taxonomy are intended primarily to provide identification, rather than phylogenetic resolution, a DNA taxonomy data base will nonetheless constitute an invaluable resource for phylogenetics. At the very least, these sequences will be sufficient for an initial hypothesis of phylogenetic relationships, in particular at the species level, even if only a single gene sequence is available. Large-scale comparisons between different parts of the tree will elucidate differences in rates and mode of molecular evolution, patterns of species diversification, variation in ecological characters, and will result in a deeper understanding of biological diversity. In addition, the very dense taxon sampling of taxonomy data bases has the potential to resolve even deep nodes, because long branches will be split up in many cases. This will provide a framework for reconstructing the Tree of Life (Box 2) and support current initiatives to establish it (NSF 02–074: http://www.nsf.gov).

The need for a new data base

The current taxonomic system is based on the Zoological and Botanical Codes of Nomenclature, which are supervised by governing bodies. In a DNA-based system, the

Box 2. Ongoing taxonomy initiatives

These websites, and the links therein, provide an excellent overview of the many ongoing initiatives to capture the current taxonomic knowledge and to enhance training and expertise in taxonomy:

- Partnership for Enhancing Expertise in Taxonomy (PEET): http://web.nhm.ukans.edu/peet/
- Integrated Taxonomic Information System (ITIS): http://www.itis.usda.gov/
- Species 2000: http://www.usa.sp2000.org/
- Convention on Biological Diversity: http://www.biodiv.org/
- Bionet International: http://www.bionet-intl.org/
- The Tree of Life Web Project: http://tolweb.org/tree/
- All Species Foundation: http://www.all-species.org/
- Global Biodiversity Information Facility: http://www.gbif.org/
- Codes of Nomenclature: http://www.biosis.org.uk/zrdocs/codes/codes.htm.

tasks for such governing bodies would change, as outlined above. In particular, the resolution of conflicts around priorities in species naming could become rather different. The current principle of collecting all previous literature to establish priority is not practical, because this literature is often only poorly accessible. DNA taxonomy would provide a chance to overhaul the current system entirely, by being directly based on accessible data base systems.

But these are not yet in existence. The current DNA data bases maintained at the National Center for Biotechnology Information (NCBI: http://www.ncbi.nlm.nih.gov/) or the European Bioinformatics Institute (EBI: http://www.ebi.ac.uk/) are not suitable for taxonomic purposes. Although they already include sequence information for well over 100 000 eukaryotic species, there is no guarantee that the correct species names were assigned by the submitter of the sequence, because there are no established taxonomic standards under which such submissions have to be done. Moreover, these data bases have no provision to include morphological, biogeographical and ecological or literature information that should be associated with a particular entry. We believe that it is unlikely, and probably even undesirable to convert these DNA data bases into ones that are taxonomically oriented. Instead, the principle of DNA taxonomy should be integrated into the current efforts for establishing universal taxonomic data bases [4,14] (Box 2).

The most urgent need would be the generation of a universal registration system for DNA taxonomy entries, akin to the accession numbers used by NCBI and EBI. In contrast to these, however, taxonomic registration numbers should not identify particular sequences, but the deposited DNA or specimen sample to which a particular sequence relates. Individual sequences from these samples could still be submitted to NCBI or EBI, but should then clearly relate to the DNA taxonomy registration number. A pilot system within such a framework has been established at the Zoologische Staatssammlung in Munich (http://www.zsm.mwn.de/DNATAX/).

High-throughput systems

DNA sequencing is often still considered to be a complex and expensive technology. It would seem that a taxonomist could identify specimens much faster and cheaper than is possible by sequencing. However, this common perception is not necessarily accurate. We live in a time where the cost of labor is rising rapidly, whereas the cost of automation keeps falling. Taxonomists take considerable time and money to train and their time is not well spent in doing routine identifications. Indeed, taxonomists tend to use their specialized knowledge to run scientific projects, but do not normally act as a service facility for the identification of specimens. Under a DNA taxonomy scheme, routine identification of specimens collected during ecological projects should be the task of high-throughput DNA-sequencing facilities. The relevant machines required for this purpose are now readily available in a price range that should enable all major museums to establish such facilities. Museums have successfully applied for technical facilities in the past, such as electron microscopy. Establishment of a DNA facility that could routinely handle ~1000 samples per day would cost approximately as much as a facility that runs a transmission and a scanning electron microscope. The material costs for each sample, including DNA extraction and sequencing of two independent regions, would be ~Euro 5 per sample in such a facility. Still, the costs for 1000 samples handled per day seems a lot by museum standards, but it is very modest when compared with ongoing genome projects. Evidently, museums could not finance this out of their already continuously declining funds, but will require entirely new funding sources. For example, one can expect that such costs can be charged to the projects that require the identification of the samples. We have no doubt that appropriate funds can be raised once a DNA-based taxonomy system finds universal acceptance. Finally, new technological developments will drive costs even lower.

The major advantage of such facilities would be twofold. First, they could provide a cost-effective service to researchers or even amateurs, who do not have direct access to such major equipment. This would also enable researchers from developing countries to directly participate in this scheme. Second, they could establish the necessary standards to ensure high-quality sequences and appropriate sample tracking systems to avoid mixups.

Conclusion

The genomic revolution of the past decade has provided us with the tools that make a universal DNA-based taxonomy system an achievable and desirable aim (Box 3). This system could help us out of the current taxonomy crisis and would give a new impetus to biodiversity research, complementing many other ongoing efforts [15] (Box 2). Most importantly, it can now be built in a way that integrates the strengths of the traditional system with the new technological possibilities. It would make full use of, and indeed requires the invaluable knowledge that has been accumulated over the centuries and it would make the knowledge of taxonomic specialists more widely

Box 3. Introducing a DNA taxonomy scheme

The following steps will be required to introduce a DNA taxonomy scheme. We believe that none of these steps should be particularly controversial and all could be agreed on relatively quickly. The eventual scheme should come out of international discussions at dedicated conferences and workshops.

Establishment of dedicated sequencing service facilities

Dedicated facilities will be required to work both on filling the data base quickly and as service facilities for all of those who want to make a taxonomic description, but do not have their own access to sequencers. Such facilities could be housed at major museums, but this is not a prerequisite. University-based or commercial solutions also seem possible.

Establishment of DNA storage facilities

This should be the task of the natural history museums, to guarantee public access. The financial requirements for this should come from extra funding, as this would constitute a new organizational task. The system should be devised such that automatic (robotic) retrieval of samples is possible. General schemes for the generation of subsamples and exact storage requirements will have to be agreed on at an international level.

Identification of pilot projects

To fill the data base quickly, the initial target projects should be those for which active research is currently being performed and funded. To make the scheme successful, expert taxonomic knowledge will be required and would have to be provided by those who run appropriate scientific projects, either in taxonomy, phylogeny or biodiversity research. Collaborations will be necessary at all scales to make the efforts worthwhile.

Governing board and curation

The commissions of nomenclature will have to make some (most probably minor) adjustments to the rules that govern nomenclature when DNA sequences are used as reference system. A governing board should be elected by the commission that would serve to integrate the different efforts and to propose the standards for the numbering system, the DNA regions to be sequenced and the procedures for storing the DNA samples. It should also propose and control rules for quality assessment and data base curation.

Data base and software development

Existing taxonomic data base schemes should be modified to integrate DNA sequences as a universal reference system. In addition, new software will be needed to enable the phylogenetic analysis of the very large number of entries that are to be expected.

accessible. It would also give natural history museums new roles as molecular facilities and guardians of biological and genomic diversity. The time has now come when molecular and morphological knowledge can and should be formally and fruitfully combined.

Acknowledgements

This article is based on discussions during a workshop on DNA taxonomy at the Zoologische Staatssammlung in Munich, Germany, financed through the Bundesministerium für Bildung und Forschung within the framework of the BIOLOG program (http://www.BIOLOG-online.info). In addition, many colleagues and three referees have contributed to the discussion and ideas.

References

1 Linnaeus, C. (1758) *Systema Naturae, Editio decima* Laur. Salvius
2 Linnaeus, C. (1753) *Species Plantarum* Laur. Salvius
3 Godfray, H.C.J. (2002) Challenges for taxonomy. *Nature* 417, 17–19
4 Bisby, F.A. *et al.* (2002) Taxonomy, at the click of a mouse. *Nature* 418, 367
5 Lee, M.S.Y. (2002) Online data base could end taxonomic anarchy. *Nature* 417, 787–788
6 Tautz, D. *et al.* (2002) DNA points the way ahead in taxonomy. *Nature* 418, 479
7 Nimis, P.L. (2001) A tale from Bioutopia. *Nature* 413, 21
8 Zietara, M.S. and Lumme, J. (2002) Speciation by host switch and adaptive radiation in a fish parasite genus *Gyrodactylus* (Monogenea Gyrodactylidae) – Appendix. *Evolution* in press
9 International Commission on Zoological Nomenclature, (1999) *International Code of Zoological Nomenclature*, (4th edn), The International Trust for Zoological Nomenclature
10 Castresana, J. (2001) Cytochrome *b* phylogeny and the taxonomy of great apes and mammals. *Mol. Biol. Evol.* 18, 465–471
11 Hebert, P.D.N. *et al.* (2002) Biological identifications through DNA barcodes. *Proc. R. Soc. Lond. Ser. B* in press
12 Pozhitkov, A.E. and Tautz, D. (2002) An algorithm and program for finding sequence specific oligo-nucleotide probes for species identification. *BMC Bioinformatics* 3, 9
13 Caterino, M.S. *et al.* (2000) The current state of insect molecular systematics: a thriving Tower of Babel. *Annu. Rev. Entomol.* 45, 1–54
14 Wilson, E.O. (2003) The encyclopedia of life. *Trends Ecol. Evol.* DOI: 10.1016/S0169-5347(02)00040-X
15 Boero, F. (2001) Light after dark: the partnership for enhancing expertise in taxonomy. *Trends Ecol. Evol.* 16, 266
16 Rossello-Mora, R. and Amann, R. (2001) The species concept for prokaryotes. *FEMS Microbiol. Rev.* 25, 39–67