ISSN 0006-2979, Biochemistry (Moscow), 2007, Vol. 72, No. 12, pp. 1377-1384. © Pleiades Publishing, Ltd., 2007.
Original Russian Text © V. S. Shneyer, 2007, published in Biokhimiya, 2007, Vol. 72, No. 12, pp. 1690-1699.

REVIEW

On the Species-Specificity of DNA: Fifty Years Later

V. S. Shneyer

Komarov Botanical Institute, Russian Academy of Sciences, ul. Prof. Popova 2, St. Petersburg, Russia, fax: (812) 346-0837; E-mail: shneyer@VS11044.spb.edu

Received July 30, 2007

Abstract—Modern approaches in DNA-based species identification are considered. Long used methods of species identification in procaryotes (G + C ratio, 16S rRNA nucleotide sequence, DNA–DNA hybridization) have recently been supplemented by the method of multilocus sequence analysis based on comparison of nucleotide sequences of fragments of several genes. Species identification in eukaryotes also employs one or two standard short fragments of the genome (known as DNA-barcodes). Potential benefits of new approaches and some difficulties during their practical realization are discussed.

DOI: 10.1134/S0006297907120127

Key words: DNA species-specificity, species identification, DNA barcoding, prokaryotes, eukaryotes

The first publications in which species-specificity of DNA was originally hypothesized used the term "species-specificity" in not entirely exact meaning; they only demonstrated some DNA difference in various rather distant species [1, 2]. These and some other studies analyzed mainly distant species of microorganisms, plants, and animals. However, species-specific taxonomic characters should describe precisely the species and allow distinguishing the species from closely related species. Nevertheless, it was clear that found correlation between DNA characters and relatedness of taxa opened wide perspectives for systematics. But only now, fifty years later, one can say that science begins the full-scale study and the use of the species specificity of DNA in the exact meaning of this term. We must note, the last is mainly attributed to eukaryotes because in microbiology the development of molecular criteria corresponding to species described earlier on the basis of morphological characters and metabolic features began earlier than in systematics of animals and plants due to poor morphological and other phenotypic features.

PROKARYOTES

In the review [3], I. N. Blokhina and G. F. Levanova already cited more than 200 references employing data on DNA structure for systematics of bacteria; and they summarized data on nucleotide base composition of DNA (proportion G + C) for more than 600 species. They also indicate that DNA composition has become a generally accepted criterion used for solution of numerous problems in systematics of bacteria, though its use cannot always guarantee distinguishing of species. Nevertheless, in some groups of bacteria even this rough parameter helped to define more precisely species composition or at least revealed necessity for perfect classification [3]. As it is known now the strains of one species usually differ by not more than 5% in GC content, whereas in various species the difference in GC content ranges from 20 to 80% [4].

Since the beginning of the 1970s, the method of molecular hybridization of DNA has been employed in systematics [5]. Soon it was found that the level of DNA–DNA hybridization is usually more than 70% among strains of the same species, whereas among different species this value is markedly lower [6, 7]. In 1987, this criterion was approved as a threshold for species demarcation and it was recommended for demarcation of species [8]. In eukaryote systematics the method of DNA hybridization is rarely used now, but it is still the main method for subdivision of bacterial strains into species and for description of new species in spite of many weaknesses and numerous critics [9, 13]. Results obtained using this method may vary due to effects of factors that are hard to take into consideration. These include physicochemical parameters, size of genome, the presence of large plasmids, purity of isolated DNA, etc. Differences between results of reciprocal reactions may reach 15% [10]. This method is time and labor consuming, rather

Abbreviations: CBOL) International Consortium Barcodes of Life; MLSA) multilocus sequence analysis; MLST) multilocus sequence typing; mtDNA) mitochondrial DNA; PCR) polymerase chain reaction.

expensive, and applicable only for cultivated strains. Finally, it is a relative measure.

Use of the 70% criterion often demonstrated an enormous amount of phenotypic diversity within one species. So as soon as methods for determination of DNA sequences were developed the search for criteria based on sequence divergence of homologous regions of bacterial genome started. A gene encoding the small subunit (16S) ribosomal RNA was chosen as the standard fragment. Data accumulated by the end of 1990s demonstrated lack of linear dependence between divergence levels of 16S rRNA and DNA–DNA hybridization in pairs of strains belonging to one species [11, 12]. Nevertheless, if the divergence level of 16S rRNA exceeded 3% for strains from various species the level of hybridization of their DNAs was always below 70% (although the converse is not always true). The threshold value of 3% was accepted for the intraspecies divergence of 16S rRNA, but it was recommended to detect the level of DNA hybridization [12]. Genome parameters of difference required to consider strains of bacteria as sufficiently divergent to be in different species (G + C content > 5%, nucleotide substitutions in 16S rDNA sequences > 3%; the level of DNA hybridization with nearest strains < 70% under standard conditions) were recommended as "minimal standards" for description of bacteria species. However, analysis of data accumulated after 1994 indicated the necessity of correction of the threshold value for similarity of 16S rRNA sequences for compared strains (at which determination of percentage of DNA hybridization is recommended) up to 98.7-99% [10].

Data on bacterial DNA hybridization values obtained during recent years are quite consistent with results obtained by comparisons of sequences of whole genomes or several genes [13, 14]. Comparison of sequences of genes common for 70 fully sequenced bacterial genomes demonstrated that for strains exhibiting more than 70% hybridization of whole-genomic DNA, the parameter of average nucleotide identity (ANI) was at least 94% [14]; for common protein-encoding genes this parameter was not less than 85% [15].

Molecular approaches, especially DNA sequencing, opened the possibility to study difficult for cultivation organisms and to determine some of their phenotypic characters [16, 17]. Surveys of sequences of bacterial genes from various environmental samples have shown that cultivable at present (and consequently more or less investigated) bacterial species represent fewer than 1% of their actual number, which may reach a million or even a billion [18, 19]. Study of 16S rRNA is widely used for analysis of bacterial diversity of various environments and for approximate assigning of non-cultivated strains to known species. Even sequencing of 16S rRNA fragments may define more precisely species belonging of phenotypically aberrant strains, which would be important in clinical practice [20].

Since resolution obtained by analysis of 16S rRNA sequences is insufficient, some attempts have been undertaken for identification of bacterial species using nucleotide sequences of more rapidly evolving protein-encoding genes. However, use of one gene usually revealed few informative sites for distinguishing related species and results can be easily distorted due to homologous recombination occurring between strains of one or even of closely related species. Although some time ago it was suggested that this happens rarely [21], recent studies have shown that in some groups of bacteria homologous recombination occurs quite frequently, and may involve several thousand nucleotides. It may occur between strains, in which divergence of nucleotide sequences reaches 25% [22], though in general the decrease in recombination probability exhibits logarithmic dependence on the increase in sequence divergence [23]. Nevertheless, it is accepted that it is more reliable to use several protein-encoding genes [13].

GenBank now contains information on more than 1200 projects on sequencing of full genomes of bacteria and archaea (about 1000 projects have been finished), and in many organisms the gene inventory has been investigated by means of the microarray analysis. Comparison of open reading frames in genomes of some pathogenic species has shown that they may have a conserved core of 50-100 genes (most of them are housekeeping genes) and a dispensable part [24]. Dispensable genes are sporadically found and their proportion in genomes of various species may significantly vary due to horizontal gene transfer and also to their partial loss [25, 26].

Multilocus sequence analysis (MLSA) has been proposed [13, 27] as an alternative to DNA hybridization method for recognizing strains belonging to the same species. With the MLSA approach it is possible to put the data obtained into a universal context even if isolates have not been cultured because this method is based on fully annotated genomes. MLSA represents the further development of multilocus sequence typing (MLST) originally designed in molecular epidemiology [28, 29] and employing DNA hybridization on microarrays (a microarray analysis). MLST distinguishes closely related strains; using this approach, it is possible to take into account allele variants differing in several variable positions. MLSA employs sequencing of several (e.g. seven) genes or their fragments (about 500 nucleotides in length) with universal primers and subsequent comparison of concatenated sequences. These should be single-copy protein-encoded genes (mainly housekeeping genes, which evolve rather slowly, but faster than 16S rRNA), which preferentially accumulate neutral substitutions. For different groups of bacteria the sets of genes may differ and may include genes of wider or narrower specificity. Species identification procedure may employ a two-step process: initial assigning e.g. to a genus may use 16S rRNA, and assigning to a species may use MLSA [13].

The demarcation of species is rather complicated in application to eukaryotes, and in the case of prokaryotes it is empirical and occasional (the species are often clusters of similar organisms based on their useful or harmful properties for man). So some microbiologists believe that species demarcation based on similarity of gene sequences of the described species (this is usual way for animals and plants species) is not promising; according to their viewpoint clusters of strains recognized by MLSA may give more reasonable subdivision of bacterial communities into species [13]. There are suggestions to create Internet databases for data obtained in various laboratories with MLSA on genes of especially interesting groups of bacteria and to develop electronic identification systems on this databases which would help to assign a strain at least to a cluster [31].

Studies of nucleotide sequences give a great body of rapidly accumulating data and knowledge for the development of various concepts of species and model for species evolution in bacteria [21, 22, 32, 33]. They are mainly based on data on homologous recombination and horizontal gene transfer as well as on limitations and barriers preventing these processes observed in bacterial populations [34]. According to modern considerations, only taxonomy based on comparison of gene sequences in combination with various phenotypic information give a consensual and pragmatic approach to defining bacterial species [31, 35].

EUKARYOTES

The first attempts to compare DNA composition of eukaryotes demonstrated "insufficiency of analysis of nucleotide base composition for investigation of species specificity and desperate need for use of deeper and finer approaches, first of all study of nucleotide sequence" [36]. Attempts to use DNA–DNA hybridization also demonstrated that due to high complexity and heterogeneity of eukaryotic genomes, this method gave phylogenetic information (requiring reasonable labor consumption) only for comparison of groups connected by a certain rather narrow level of relatedness.

Effective study of very closely related groups became possible only after the development of methods for more specific comparison of nucleotide sequences: initially the method of analysis of restriction fragment length polymorphism (RFLP) and later, after the development of PCR, the method of random amplification of polymorphic DNA (RAPD), analysis of microsatellites, of inter simple sequence repeats (ISSR)-PCR, etc. [37-39]. Indeed, these methods could reveal species-specific differences and even intraspecies polymorphism. These methods are still widely used especially in studies of plants. But only the development of methods for sequencing made possible direct comparison of gene sequences and search among them those sufficiently variable to serve as the species-specific markers.

Phylogenetic studies of animals at the species level initially employed mitochondrial genes, which exhibit higher polymorphism than nuclear genes; this gave the possibility to compare and distinguish closely related species. Mitochondrial DNA can be easily isolated (especially from damaged material), it is inherited via the maternal line and is less subjected to recombination than nuclear DNA. The most frequently studied genes are the genes encoding cytochrome *b* (*cytb*), subunits of cytochrome *c* oxidase (*cox1* and *cox2*), and so-called "control region" or D-loop. Other genes are also used. These include genes encoding subunits of nicotinamide dinucleotide dehydrogenase (ND1, ND2, ND3, ND4, ND5), 12S and 16S rRNA, and other fragments of the mitochondrial genome. In nuclear genomes, rRNA internal transcribed spacers ITS1-2, histone H3, transcription elongation factor 1-A (*tef*), β-tubulin (*tub*), and the most recently recombination activation gene (*rag*), introns of actin gene, myoglobin, β-fibrinogen genes, and some others were analyzed. Plant phylogeny studies employed non-coding regions of the chloroplast genome: introns of genes *rpl16*, *rps16*, *rpoC1*, *trnK*, *trnL* and intergene spacers (*trnL-trnF*, *trnT-trnL*, *atpB-rbcL*, *psbA-trnH*). The inventory of nuclear regions was rather limited: rRNA internal transcribed spacers ITS1-2 and rarely introns of RNA polymerases, genes of alcohol dehydrogenases (*Adh*), and some others were preferentially used.

Advances in technology and lowering costs of DNA sequencing resulted in sequencing of some genome regions in several thousand species; for example, *cytb* and *cox 1* in animals, ITS1-2 and intron of *trnL* gene in plants [40, 41]. Recently, gene sequencing became available not only for single specimen from species or intraspecies taxa, but also for more or less representative sample. This stimulated study of species at a new level. For some animals species (agricultural animals and birds, marketable fishes, strictly protected species) species identification systems for detection of corresponding DNA in foodstuff, smuggling products and also for forensic medicine have already been developed; these are based on *cytb*, and also 12S and 16S rRNA and microsatellites [42]. Special computer identification system "DNA Surveillance" has been developed for the order Cetacea, consisting of 80 species [43]. It is possible that they are the only taxonomic group of this rank with a species list studied so fully with two mitochondrial markers (*cytb* and control region). The representative set of reference sequences derives from expertly identified specimens.

Selection of gene by scientists depended on accidental factors; for example, gene *cytb* and the control region of mtDNA are most frequently used for phylogenetic studies of vertebrates at species level, but *cytb* has rarely been used for investigation of such huge and diverse group as insects in spite of good resolution received with this

region [44, 45]. Use of different molecular markers by various authors complicates analysis and summarization of results obtained in different laboratories; this suggests the necessity for coordinated studies.

Special attention [46] is given to the gene encoding cytochrome *c* oxidase subunit (*cox1*), presented in all studied mitochondrial genomes and actively studied for more than 20 years. The gene consists of 1540 nucleotides. Phylogenetic studies frequently employ its variable 5′-half of about 650 nucleotides in length; it is positioned between two conserved sequences. Using standard primers constructed to these sequences, it is possible to amplify corresponding DNA fragment in most multicellular organisms [47].

Comparison of sequences of 5′-fragments of *cox1* from specimens belonging to various species of birds, butterflies, spiders and other groups of animals obtained by the authors or taken from GenBank has shown [46, 48-50] that interspecies divergence of sequences is >2-3% (6-23%, average 11.3%), whereas intraspecies one is usually <3%, frequently 1% (divergence of sequences was determined by calculating distances using the Kimura two parametric model). The values of intra- and interspecies divergence of 5′-fragments of *cox1* differed by 5-20-fold. As the result of these comparisons the suggestion was proposed to use the 5′-fragments of *cox1* as a molecular barcode for identification of animal species and to set a rule of 10-fold differences (10× SST, species-screening threshold) in values of intraspecies and interspecies divergence of *cox1*-fragment sequences or employ the 2-3% sequence divergence as the species threshold. The second criterion includes the existence of reciprocal monophyly, representing lack of sequence overlapping [49].

This represented a good background for the new international program "Barcode of Life Initiative" proposed for studies and molecular cataloging of species diversity of all animals and plant world of the Earth. The major goal of this program is to provide molecular identification of organisms using standardized DNA region (DNA barcode) and to create a special database that would be more taxonomically accurate and would have more rigorous rules for entry of the data compared with existing GenBank databases. In the opinion of the authors of this initiative DNA barcoding, besides rapid and reliable assigning specimens to known species would really facilitate discovery of new species through identification of lineages divergent by the barcode as well as specification of interspecies borders. Molecular identification is important both for science (systematics, ecology, biogeography, etc.) and practice (protection of environment, monitoring, quarantine services, medicine, veterinary, forensics, control of medicinal materials and products, etc.). This program attracts special interest of specialists studying groups of organisms, where traditionally used approaches are rather ineffective and where it is impossible to identify a specimen by its appearance and special (time consuming) procedures are thus required; this is also important for cryptic (morphologically similar) species and/or very small organisms, such as sponges, nematodes, flat worms, copepods, insects, spiders, algae, mosses, and many others. It is very important that for the analysis of DNA region it is sufficient to have a tiny fragment of any tissue of an organism at any stage of its development including preserved collection specimens.

For coordination of studies and database development the system known as BOLD (The Barcode of Life Data Systems; www.barcodinglife.org) and the international consortium CBOL (Consortium for the Barcode of Life) with secretariat at the Natural History Museum (Washington) have been organized. This consortium joins together all institutions and persons working in the stream of the approach. According to the standards developed by CBOL, DNA sequences selected as the standard should meet the following criteria: i) they should be rather short (not more than 700-800 nucleotides in length for facilitation and economy of isolation, amplification and sequencing); ii) they should be rather variable (for discrimination of closely related species) but flanked with conserved regions required for amplifications with primers of wide specificity; iii) they should be easily aligned (i.e. they should contain very few indels). At the moment only one region of DNA, a fragment of *cox1*, corresponding to the fragment of mitochondrial gene of mouse DNA (positions from 58 to 705 of 648 bp in length), has been approved as a barcode. A sequence added to the database as the barcode should include not less than 500 neighboring nucleotides, should be read in both directions, should include sequences of forward and reverse primers, and should not contain more than 1% of polymorphic positions. In the case of assembly of the sequence from several amplicons using several primers, all information should be presented with indications of primer nomenclature.

Potential applicability of the *cox1* fragment as the DNA barcode has been demonstrated for representatives of many groups of animals, possibly for some algae and fungi, but definitely not for all eukaryotes. Many authors doubt that just one region of the genome can serve as a reliable species marker, especially that it would be possible to select one DNA region for identification of all animal species [51-53]. In some animals *cox1* gene variability is low: for example in sponges [54] and cnidarians [55] for which low variability of mitochondrial genes including *cytb* [56] has been demonstrated earlier. Other organisms are characterized by too high variability of *cox1*, and in this case it is impossible to use conserved primers (e.g. in copepods [57], amphibians [53], for which more suitable mitochondrial genes encoding ND1 [58] or 16S rRNA [59] may be used). In some animals *cox1* is subjected to large rearrangements and therefore it is impossible to use universal primers (e.g. nematodes, flukes, tardigrades [60]). Such groups require the development of specific

primers or the use of other regions of the genome. Alternative markers for many organisms (especially invertebrates) may include fragments of nuclear ribosomal 18S (SSU) and 28S (LSU) rRNA [61, 62], representing mosaics of alternating conserved and variable regions, which make possible to use universal primers for their amplification.

It has also been proposed to increase the length of the analyzed fragment in the case of necessity (modern technologies provide reasonable opportunities to do that [63]). For example, in sponges (the group characterized by difficulty of species identification even for experienced taxonomists) acceptable level of variability may be achieved by use of a longer (approximately by 460 nucleotides) fragment of *cox1*, rather than of the recommended standard region [54].

Fungal mitochondrial genomes still require better investigation, but it is known that the length of *cox 1* in them is subjected to significant variations due to the presence of introns, and so in phylogenetics of fungi it is not usually employed. Studies of closely related fungal species preferentially employed fragments of nuclear DNA encoding β-tubulin (*tub*), rRNA ITS1 and ITS2, translation elongation factor 1-A (*tef*), and variability of each fragment used separately was insufficient for acceptable resolution [64]. Special analysis of GenBank data on fungal mitochondrial genomes has shown the uneven distribution of *cox1* introns including the region corresponding to the barcode [65]. Promising results have been obtained in the barcode testing of the *cox1* fragment of ascomycetes taxa from one of the most important for man representatives, *Penicillium* genus [65]. (This is a taxonomically complex group for which the identification based on molecular markers has not been developed [65].) The variability of this fragment was satisfactory and introns were found in minor group of strains, but in that case it would be possible to use reverse transcription (RT)-PCR.

For plants a DNA fragment candidate for barcode has not been found yet. In most cases, the mitochondrial DNA genes including *cox1* are inapplicable due to low and uneven variability and frequent large structural rearrangements [66, 67]; some algae possibly represent an exception. Identification employing *cox1* fragment was successful in some groups of red algae [68]. This was promising. However, attempts to use this fragment for analysis of complex groups of species of brown algae have shown lack of correlation between the haplotype distribution and data obtained by means of two other markers, nuclear ITS and the plastid spacer *rbcSp*, frequently used by algologists [69]. Thus, potential species specificity of *cox1* in algae is still questioned.

Variability of chloroplast-encoding genes (*rbcL*, *ndhF*, *matK*, *atpB*) as well as non-coding introns and spacers was often insufficiently informative [70]. Analyses *in silico*, appearing in the process of accumulation of data on full-sized chloroplast genomes, demonstrate the presence of highly variable regions that have not been used yet in phylogenetic studies [71, 72]. However, suitable candidate fragments among them, as well as among low-copy nuclear genes, which would be used as a barcode, have not also been found yet [73, 74]. Now the chloroplast spacer *psbA-trnH* is considered as a putative barcode candidate [75, 76], but recently it has been demonstrated that its variability is insufficient for phylogeny studies at the species level [77, 78].

In plants the most variable are nuclear spacers ITS, and they are more frequently used due to several additional advantages: biparental inheritance, development of many primers, with which they are readily amplified (including herbarium samples); however, some problems arise during their use [79, 80]. The most frequent are indels and possible existence of multiple non-identical paralogous copies, including nonfunctional pseudogenes (sometimes within one genome); they cannot be always found during direct sequencing without cloning, and cloning significantly complicates the work and increases expenses. On the other hand, successful phylogenetic analysis of closely related species by means of ITS pseudogenes free of constraints during accumulation of mutations has recently been reported, and these are more variable than corresponding functional paralogs [81].

It should be noted that the problem of pseudogenes complicates use of the mitochondrial gene as well. In many animals and plants there are nuclear copies of mitochondrial genes (numt), which represent (in most cases) nonfunctional genes with nucleotide sequences markedly differing from the corresponding mitochondrial sequences. These numts are characterized by uneven distribution, and some groups and even closely related species may differ in their presence and abundance. Since nuclear genes evolve more slowly than mitochondrial genes (in animals) they will exhibit preferential amplification with universal primers. The presence of unfound numt in the analyzed sample will increase the detected level of divergence, and therefore it is important to monitor the presence of pseudogenes [82].

Some authors believe that nuclear ITS are still the best choice for studies of plants (possibly together with chloroplast, mitochondrial, or low-copy nuclear regions) because of their known and well-studied shortcomings and constraints [83, 84]. It is also possible to develop two- or three-step DNA barcoding of plants: the first step would employ a conserved region of the genome (e.g. *rbcL*) which may refer a testing specimen say to a family, whereas the second step would use one of more specific markers [85, 86]. However, evaluation of applicability of a putative marker as the species-specific indicator is mainly based on the information about its resolution for phylogeny reconstruction. But these tasks are not identical. In the phylogenetic analysis evaluation of similarity (distance) usually involves all or informative positions (sometimes including indels). Identification may be car-

ried out by determining species-specific indels, autapomorphies, or combinations of rare substitutions. It was also found that sometimes DNA regions with variability insufficient for phylogeny reconstruction might provide reliable species identification after optimization of criteria [83, 85].

Now wide scale testing of the barcode *cox1* employs various taxonomic groups. During the past two-three years many studies have been carried out and in half of them the values of interspecies divergence of the *cox1* fragment are within the proposed 3%, and there is no overlapping between intra- and interspecies variability and most species have been reliably identified.

However, other studies have demonstrated overlapping both within certain groups [53, 87] and during *in silico* analyses covering larger groups [88, 89]. In some cases such overlapping could be interpreted as inapplicability of certain genome region as a barcode whereas in others this could demonstrate imperfectness of the taxonomical system analyzed and superfluous splitting of species. Thus, it is reasonable to conclude that taxonomically well-understood groups as well as reference databases based on representative samples for all the species covering intraspecific morphological, geographical, and ecological diversity of species are necessary prerequisites for successful use of DNA barcodes as the species identification system. This is especially important for young recently diverging groups for which DNA barcoding may be ineffective. Such species may undergo frequent hybridization between each other and constitute species complexes. Multiple haplotypes unevenly distributed among individuals have been found within these complexes, and analyses of different combination of single exemplars could have resulted in a different inferred phylogeny [90, 91]. In the complexes of species with frequent intertaxa hybridization, distribution of molecular markers frequently reflects geographic distribution of groups due to introgression [92, 93].

Studies of four fragments of the genome (one nuclear and three mitochondrial including *cox1*, of total lengths of 2000 nucleotides) of the complex of beetles of genus *Copelatus* from Fiji Islands have shown marked diversity of haplotypes forming a continuum of related groups [94]. In other cases studies of molecular variability of one gene revealed several lineages corresponding to cryptic species within one morphologically non-polymorphic genus. For example, *cox1* variability in the *Astraptes fulgerator* butterflies (about 500 individuals were studied) revealed 10 lineages corresponding (according to the authors' viewpoint) to species and this correlated with differences in food preferences and in color of their larvae [95]. Variability of mitochondrial 16S rRNA suggested the presence of 15 cryptic species in the *Schindleria praematura* fish [96]. Based on distribution of molecular lineages pilot systems for subsequent detailed taxonomic analysis can be created. It should be noted that the idea of using DNA barcodes for identification of new species has been criticized from many viewpoints [97, 98], including concerns that this approach does not take into consideration multiple concepts and criteria elaborated for species and related uncertainty of this definition, which still exists. However, one cannot deny that use of molecular markers of rather narrow specificity allows to better understand not only structure of biodiversity groups, but also taxonomic importance and the mode of evolution of many other characters.

Now active search and analysis of species-specific DNA regions is carried out and areas of their applications are extended as well. There are several tens of pilot projects on DNA barcoding of butterflies, birds, fishes, invading species, etc., preferentially based on good collections available. Newly published results are immediately analyzed and criticized by colleagues and soon it becomes clear in which cases and in which groups DNA barcoding is effective. Meanwhile CBOL already plans to build a library of reference DNA barcode sequences for all known eukaryotic species and to develop a portable field identifier, which would analyze DNA barcodes in a small fragment of tissues from a particular organism and determine species belonging by search in a database.

E. Chargaff, the author who was the first to report on the differences in DNA composition in various species [1], wrote many years later that his paper carried the seeds for the future [99]. Such seeds for the future were also in the papers on species-specificity of DNA [2, 36], this can be demonstrated by advances and also by perspectives in this field.

This work was supported by the Russian Foundation for Basic Research (grant 06-04-48399) and the Program "The genofond dynamics".

REFERENCES

1. Chargaff, E. (1950) *Experientia*, **6**, 201-209.
2. Spirin, A. S., Belozersky, A. N., Shugaeva, N. V., and Vanyushin, B. F. (1957) *Biokhimiya*, **22**, 744-754.
3. Blokhina, I. N., and Levanova, G. F. (1972) *DNA Structure and Position of Organisms in the System* (Belozersky, A. N., and Antonov, A. S., eds.) [in Russian], Moscow University Press, Moscow, pp. 91-134.
4. Muto, A., and Osawa, S. (1987) *Proc. Natl. Acad. Sci. USA*, **84**, 166-169.
5. De Ley, J., Cattoir, H., and Reynaerts, A. (1970) *Eur. J. Biochem.*, **12**, 133-142.
6. Johnson, J. (1973) *Int. J. Syst. Bacteriol.*, **23**, 308-315.
7. Tourova, T. P., and Antonov, A. S. (1987) *Meth. Microbiol.*, **19**, 333-355.
8. Wayne, L. G., Brenner, D. J., Colwell, R. R., Grimont, P. A. D., Kandler, O., Krichevsky, M. I., Moore, W. E. C., Murray, R. G. E., Stackebrandt, E., Starr, M. P., and Truper, H. G. (1987) *Int. J. Syst. Bacteriol.*, **37**, 463-464.

9. Konstantinidis, K. T., Ramette, A., and Tiedje, J. M. (2006) *Appl. Environ. Microbiol.*, **72**, 7286-7293.
10. Stackebrandt, E., and Ebers, J. (2006) *Microbiol. Today*, **33**, 152-155.
11. Fox, G. E., Wisotzkey, J. D., and Jurtshuk, P. (1992) *Int. J. Syst. Bacteriol.*, **42**, 166-170.
12. Stackebrandt, E., and Goebel, B. M. (1994) *Int. J. Syst. Bacteriol.*, **44**, 846-849.
13. Gevers, D., Cohan, F. M., Lawrence, J. G., Spratt, B. G., Coenye, T., Feil, E. J., Stackerbrandt, E., van de Peer, Y., Vandamme, P., Thompson, F. L., and Swing, J. (2005) *Nat. Rev. Microbiol.*, **3**, 733-739.
14. Konstantinidis, K. T., and Tiedje, J. M. (2005) *Proc. Natl. Acad. Sci. USA*, **102**, 2567-2572.
15. Goris, J., Konstantinidis, K. T., Klappenbach, J. A., Coenye, T., Vandamme, P., and Tiedje, J. M. (2007) *Int. J. Syst. Evol. Microbiol.*, **57**, 81-91.
16. Ward, D. M., Weller, R., and Bateson, M. M. (1990) *Nature*, **345**, 63-65.
17. Torsvik, V., Daae, F. L., Sandaa, R. A., and Ovreas, L. (1998) *J. Biotechnol.*, **64**, 53-62.
18. Curtis, T. P., and Sloan, W. T. (2004) *Curr. Opin. Microbiol.*, **7**, 221-226.
19. Giovannoni, S. J., and Stingl, U. (2005) *Nature*, **437**, 343-348.
20. Petti, C. A., Polage, C. R., and Schreckenberger, P. (2005) *J. Clin. Microbiol.*, **43**, 6123-6125.
21. Cohan, F. M. (2001) *Syst. Biol.*, **50**, 513-524.
22. Vulic, M., Dionisio, F., Taddei, F., and Radman, M. (1997) *Proc. Natl. Acad. Sci. USA*, **94**, 9763-9767.
23. Majewski, J., Zawadzki, P., Pickerill, P., Cohan, F. M., and Dowson, C. G. (2000) *J. Bacteriol.*, **182**, 1016-1023.
24. Ochman, H., and Jones, I. B. (2000) *EMBO J.*, **19**, 6637-6643.
25. Welch, R. A., Burland, V., Plunkett, G., Redford, P., Roesch, P., Rasko, D., Buckles, E. L., Liou, S.-R., Boutin, A., Hackett, J., Stroud, D., Mayhew, G. F., Rose, D. J., Zhou, S., Schwartz, D. C., Perna, N. T., Mobley, H. L. T., Donnenberg, M. S., and Blattner F. R. (2002) *Proc. Natl. Acad. Sci. USA*, **99**, 17020-17024.
26. Ochman, H., and Santos, S. R. (2005) *Infect. Genet. Evol.*, **5**, 103-108.
27. Hanage, W. P., Kaijalainen, T., Herva, E., Saukkorupi, A., Syrjanen, R., and Spratt, B. G. (2005) *J. Bacteriol.*, **187**, 6223-6230.
28. Maiden, M. C., Bygraves, J. A., Feil, E., Morelli, G., Russell, J. E., Urwin, R., Zhang, Q., Zhou, J., Zurth, K., Caugant, D. A., Feavers, I. M., Achtman, M., and Spratt, B. G. (1998) *Proc. Natl. Acad. Sci. USA*, **95**, 3140-3145.
29. Platonov, A. E., Shipulin, G. A., and Platonova, O. V. (2000) *Russ. J. Genet.*, **36**, 481-487.
30. Cohan, F. M., and Perry, E. B. (2007) *Curr. Biol.*, **17**, R373-R386.
31. Hanage, W. P., Fraser, C., and Spratt, B. G. (2006) *Philos. Trans. R. Soc. London B*, **361**, 1917-1927.
32. Dykhuizen, D. E., and Green, L. (1991) *J. Bacteriol.*, **173**, 7257-7268.
33. Staley, J. T. (2006) *Philos. Trans. R. Soc. London B*, **361**, 1899-1909.
34. Ochman, H., and Davalos, L. M. (2006) *Science*, **311**, 1730-1733.
35. Ochman, H., Lerat, E., and Daubin, V. (2005) *Proc. Natl. Acad. Sci. USA*, **102**, 6595-6599.
36. Belozersky, A. N., and Spirin, A. S. (1960) *Izv. Akad Nauk SSSR, Biol. Ser.*, No. 1, 64-81.
37. Shneer, V. S. (1991) *Bot. Zh.*, **76**, 17-32.
38. Bannikova, A. A. (2004) *Zh. Obshch. Biol.*, **65**, 278-305.
39. Sulimova, G. E. (2004) *Usp. Sovrem. Biol.*, **124**, 260-271.
40. Nickrent, D. L., Garcia, M. A., Martin, M. P., and Mathiasen, R. L. (2004) *Amer. J. Bot.*, **91**, 125-138.
41. Hajibabaei, M., Singer, G. A. C., Hebert, P. D. H., and Hickey, D. A. (2007) *Trends Genet.*, **23**, 167-172.
42. Teletchea, F., Maudet, C., and Hanni, C. (2005) *Trends Biotechnol.*, **23**, 359-366.
43. Ross, H. A., Lento, G. M., Dalebout, M. L., Goode, M., Ewing, G., McLaren, P., Rodrigo, A. G., Lavery, S., and Baker, C. S. (2003) *J. Hered.*, **94**, 111-114.
44. Simmons, R. B., and Weller, S. J. (2001) *Mol. Phyl. Evol.*, **20**, 196-210.
45. Caterino, M. S., Cho, S., and Sperling, F. A. H. (2000) *Annu. Rev. Entomol.*, **45**, 1-54.
46. Hebert, P. D. N., Cywinska, A., Ball, S. L., and de Waard, J. R. (2003) *Proc. R. Soc. London B*, **270**, 313-321.
47. Folmer, O., Black, M., Hoeh, W., Lutz, R., and Vrigenhoek, R. (1994) *Mol. Marine Biol. Biotechnol.*, **3**, 294-299.
48. Hebert, P. D. N., Penton, E. H., Burns, J. M., Janzen, D. H., and Hallwachs, W. (2004) *Proc. Natl. Acad. Sci. USA*, **101**, 14812-14817.
49. Hebert, P. D. N., Stoeckle, M. Y., Zemlak, T. S., and Francis, C. M. (2004) *PLoS Biol.*, **2**, 1657-1663.
50. Barrett, R. D. H., and Hebert, P. D. N. (2004) *Can. J. Zool.*, **83**, 481-491.
51. Dalebout, M. L., Baker, C. S., Mead, J. G., Cockcroft, V. G., and Yamada, T. K. (2004) *J. Hered.*, **95**, 459-473.
52. Armstrong, K. F., and Ball, S. L. (2005) *Philos. Trans. R. Soc. London B*, **360**, 1813-1823.
53. Vences, M., Thomas, M., Bonett, R. M., and Vieites, D. R. (2005) *Philos. Trans. R. Soc. London B*, **360**, 1859-1868.
54. Erpenbeck, D., Hooper, J. N. A., and Worheide, G. (2006) *Mol. Ecol. Notes*, **6**, 550-553.
55. Hellberg, M. E. (2006) *BMC Evolut. Biol.*, **6**, e24.
56. Van Oppen, M. J. H., Willis, B. L., and Miller, D. J. (1999) *Proc. R. Soc. London B*, **266**, 179-183.
57. Goetze, E. (2003) *Proc. R. Soc. London B*, **270**, 2321-2331.
58. Camargo, A., De Sa, R. O., and Heyer, W. R. (2006) *Biol. J. Linn. Soc.*, **87**, 325-341.
59. Vences, M., Thomas, M., van der Meijden, A., Chiari, Y., and Vieites, D. R. (2005) *Front. Zool.*, **2**, e5.
60. Blaxter, M., Mann, J., Chapman, T., Thomas, F., Whitton, C., Floyd, R., and Abebe, E. (2005) *Philos. Trans. R. Soc. London B*, **360**, 1935-1943.
61. Floyd, R., Eyualem, A., Papert, A., and Blaxter, M. (2002) *Molec. Ecol.*, **11**, 839-850.
62. Markmann, M., and Tautz, D. (2005) *Philos. Trans. R. Soc. London B*, **360**, 1917-1924.
63. Roe, A. D., and Sperling, F. A. H. (2007) *Mol. Phyl. Evol.*, **44**, 325-345.
64. Kauserud, H., Svegarden, I. B., Decock, C., and Hallenberg, N. (2007) *Mol. Ecol.*, **16**, 389-399.
65. Seifert, K. A., Samson, R. A., deWaard, J. R., Houbraken, J., Levesque, C. A., Moncalvo, J.-M., Louis-Seize, G., and

Hebert, P. D. N. (2007) *Proc. Natl. Acad. Sci. USA*, **104**, 3901-3906.
66. Wolfe, K. H., Li, W.-H., and Sharp, P. M. (1987) *Proc. Natl. Acad. Sci. USA*, **84**, 9054-9058.
67. Cho, Y., Mower, J. P., Qui, Y.-L., and Palmer, J. D. (2004) *Proc. Natl. Acad. Sci. USA*, **101**, 17741-17746.
68. Saunders, G. W. (2005) *Philos. Trans. R. Soc. London B*, **360**, 1879-1888.
69. Lane, C. E., Lindstrom, S. C., and Saunders, G. W. (2007) *Mol. Phylogen. Evol.*, **44**, 634-648.
70. Small, R. L., Ryburn, J. A., Cronn, R. C., Seelanan, T., and Wendel, J. F. (1998) *Amer. J. Bot.*, **85**, 1301-1315.
71. Show, J., Lickey, E. B., Beck, J. T., Farmer, S. B., Liu, W., Miller, J., Chow, S. K., Winder, C. T., Schilling, E. E., and Small, R. L. (2005) *Amer. J. Bot.*, **92**, 142-166.
72. Shaw, J., Lickey, E. B., Schilling, E. E., and Small, R. L. (2007) *Amer. J. Bot.*, **94**, 275-288.
73. Mort, M. E., Archibald, J. K., Randle, C. P., Levsen, N. D., O'Leary, T. R., Topalov, K., Wiegand, C. M., and Crawford, D. J. (2007) *Amer. J. Bot.*, **94**, 173-183.
74. Hughes, C. E., Eastwood, R. J., and Bailey, C. D. (2006) *Philos. Trans. R. Soc. London B*, **361**, 211-225.
75. Kress, W. J., Wurdack, K. J., Zimmer, E. A., Weigt, L. A., and Janzen, D. H. (2005) *Proc. Natl. Acad. Sci. USA*, **102**, 8369-8374.
76. Kress, W. J., and Erickson, D. L. (2007) *PLoS Biol.*, **6**, e508.
77. Kim, S. C., Crawford, D. J., Jansen, R. K., and Santos-Guerra, A. (1999) *Plant Syst. Evol.*, **215**, 85-99.
78. Peterson, A., John, H., Koch, E., and Peterson, J. (2004) *Plant. Syst. Evol.*, **245**, 145-162.
79. Alvarez, I., and Wendel, J. F. (2003) *Mol. Phylogen. Evol.*, **29**, 417-434.
80. Bailey, C. D., Carr, T. G., Harris, S. A., and Hughes, C. E. (2003) *Mol. Phylogen. Evol.*, **29**, 435-455.
81. Ochieng, J. W., Henry, R. J., Baverstock, P. R., Steane, D. A., and Shepherd, M. (2007) *Mol. Phylogenet. Evol.*, **44**, 752-764.
82. Bensasson, D., Zhang, D.-X., Hartl, D. L., and Hewitt, G. M. (2001) *Trends Ecol. Evol.*, **16**, 314-321.
83. Gemeinholzer, B., Oberprieler, C., and Bachmann, K. (2006) *Taxon*, **55**, 173-187.
84. Feliner, G. N., and Rossello, J. A. (2007) *Mol. Phylogenet. Evol.*, **44**, 911-919.
85. Chase, M. W., Salamin, N., Wilkinson, M., Dunwell, J. M., Kesanakurthi, R. P., Haidar, N., and Savolainen, V. (2005) *Philos. Trans. R. Soc. London B*, **360**, 1889-1895.
86. Newmaster, S. G., Fazecas, A. J., and Ragupathy, S. (2006) *Can. J. Bot.*, **84**, 335-341.
87. Meyer, C. P., and Paulay, G. (2005) *PloS Biol.*, **3**, e422.
88. Meier, R., Shiyang, K., Vaidya, G., and Ng, P. K. L. (2006) *Syst. Biol.*, **55**, 715-728.
89. Cognato, A. I. (2006) *J. Econom. Entomol.*, **99**, 1037-1045.
90. Shaw, J., and Small, R. L. (2004) *Amer. J. Bot.*, **91**, 985-996.
91. Shaw, J., and Small, R. L. (2005) *Amer. J. Bot.*, **92**, 2011-2030.
92. Matos, J. A., and Schaal, B. A. (2000) *Evolution*, **54**, 1218-1233.
93. Funk, D. J., and Omland, K. E. (2003) *Annu. Rev. Ecol. Evol. Syst.*, **34**, 397-423.
94. Monaghan, M. T., Balke, M., Pons, J., and Vogler, A. P. (2006) *Proc. Roy. Soc. Lond. B*, **273**, 887-893.
95. Hebert, P. D. N., Penton, E. H., Burns, J. M., Janzen, D. H., and Hallwachs, W. (2004) *Proc. Natl. Acad. Sci. USA*, **101**, 14812-14817.
96. Kon, T., Yoshino, T., Mukai, T., and Nishida, M. (2007) *Mol. Phylogen. Evol.*, **44**, 53-62.
97. Wheeler, Q. D. (2004) *Philos. Trans. R. Soc. London B*, **359**, 571-583.
98. Rubinoff, D., Cameron, S., and Will, K. (2006) *J. Hered.*, **97**, 581-594.
99. Chargaff, E. (1979) *Ann. New York Acad. Sci.*, **325**, 345-360.