Sequence analysis

Ribosomal RNA as molecular barcodes: a simple correlation analysis without sequence alignment

K. H. Chu1,*, C. P. Li1 and J. Qi2
1Department of Biology, The Chinese University of Hong Kong, Hong Kong, China and 2Center for Comparative Genomics and Bioinformatics, Pennsylvania State University, University Park, PA 16802, USA

Received on January 6, 2006; revised and accepted on April 11, 2006
Advance Access publication April 13, 2006
Associate Editor: Chris Stoeckert

ABSTRACT

Motivation: We explored the feasibility of using unaligned rRNA gene sequences as DNA barcodes, based on correlation analysis of composition vectors (CVs) derived from nucleotide strings. We tested this method with seven rRNA (including 12, 16, 18, 26 and 28S) datasets from a wide variety of organisms (from archaea to tetrapods) at taxonomic levels ranging from class to species.

Result: Our results indicate that grouping of taxa based on CV analysis is always in good agreement with the phylogenetic trees generated by traditional approaches, although in some cases the relationships among the higher systemic groups may differ. The effectiveness of our analysis might be related to the length and divergence among sequences in a dataset. Nevertheless, the correct grouping of sequences and accurate assignment of unknown taxa make our analysis a reliable and convenient approach in analyzing unaligned sequence datasets of various rRNAs for barcoding purposes.

Availability: The newly designed software (CVTree 1.0) is publicly available at the Composition Vector Tree (CVTree) web server http://cvtree.cbi.pku.edu.cn

Contact: kahouchu@cuhk.edu.hk

INTRODUCTION

It is estimated that there are between 3.6 million and 100 million species on earth (Heywood and Watson, 1995), which are valuable biological resource for the human civilization. In the recent decade, the loss of biodiversity has been recognized as a major global environmental problem and much effort has been targeted on biodiversity conservation. Yet a major obstacle in accessing human impact on the biosphere is what has often been referred to as ‘taxonomic impediment’, which reflects the lack of taxonomic expertise in many groups of living organisms, and information in taxonomy is not always accessible and intelligible to biologists who are not taxonomists (Minelli, 2003). To overcome this problem, genetic information, specifically DNA sequences, has been suggested to serve as a criterion, or at least a complement, in taxonomic identification (e.g. Blaxter, 2003; Mallet and Willmott, 2003; Tautz et al., 2003; Savolainen et al., 2005).

Hebert et al. (2003a, b) proposed the use of mitochondrial cytochrome c oxidase subunit I (COI) as ‘DNA barcodes’ for species identification in the animal kingdom and its applicability has been demonstrated in a wide variety of organisms (e.g. Hebert et al., 2004a,b; Saunders, 2005; Ward et al., 2005). Yet closely related species may have identical or nearly identical COI sequences (Harrison, 2004; Lorenz et al., 2005; Zhang et al., 2005). It has also been suggested that it is undesirable to rely on a single sequence for taxonomic identification (Sites and Crandall, 2000; Noé and Kucherov, 2004). Since base insertions and deletions (indels) are common in RNA sequences, every sequence with indels has to be assigned gaps for alignment with the others. Since no universal alignment parameters are defined, assigning gaps into DNA sequence is subjective (Geiger, 2002), and there is no consensus on what defines a ‘good’ or a ‘best’ multiple alignment (Wheeler, 1996). As a result, even when the alignment process is performed carefully by experienced researchers, human errors can be introduced, particularly in some RNA sequences for which no closely related sequences are available for use as reference. Besides ambiguity in multiple sequence alignment, this process often has to be repeated whenever a new sequence is added to a dataset before analysis. It is estimated that 200,000 barcode records will be added into the database each year (Hajibabaei et al., 2005). With such a large dataset, sequence alignment in the barcode project would become tedious and time consuming. While it may be argued that sequences in a DNA barcode database can be divided to subsets (each representing an appropriate taxonomic level such as class, order or family) which can be aligned independently for analysis, each of these datasets would still make up of hundreds to thousands of sequences which are massive for the alignment procedure. For instance, the fish family Cichilidae of about 1300 described species (Kullander, 1998) and the insect family Tipulidae (craneflies) of up to 15,000 species (Alexander, 1920) are examples of taxa that are preferably to be analyzed as a group. Similarly, there are about 22,000 described species of nematodes and new DNA sequences of this group (for which the corresponding descriptive taxonomy may not be available) possibly have to be analyzed with those from all nematodes as a whole. Another drawback of the alignment
procedure is that ideally, the DNA sequences for barcoding purpose should be intact, i.e. not incorporated with any artifacts, including gaps. Otherwise, the same sequence may be referred to as different should be intact, i.e. not incorporated with any artifacts, including gaps. Thus, sequence alignment is a major obstacle that limits the effective use of rRNAs for barcode purposes.

In this present study, we attempted to analyze rRNA sequences without alignment, using a simple correlation analysis based on composition vectors (CVs) derived from sequence data (Yu and Jiang, 2001; Chu et al., 2004; Qi et al., 2004a,b), with a view to test the feasibility of using rRNAs as molecular barcodes. In line with the studies which demonstrate the use of COI as DNA barcodes (Hebert et al., 2003a,b), we have considered to assemble large datasets of rDNA sequences from GenBank database for our feasibility study. Yet this approach needs construction of trees based on alignment for comparative purpose, which would involve ambiguity of alignment as well as subjectivity in the choice of tree construction methods. Thus, we have taken an alternative strategy by comparing our approach with published rRNA trees in the literature, plus an unpublished tree from our own research. We analyzed a total of seven rRNA datasets from a wide variety of organisms and taxonomic levels, from archaea to tetrapods, from class to species. The results demonstrated that unaligned rRNA gene sequences could be used as convenient and reliable DNA barcodes.

### MATERIALS AND METHODS

Sequences from six published rRNA datasets (Aralah et al., 1996; Ro et al., 1997; de Bellocq et al., 2001; Shull et al., 2001; Rickard et al., 2002; Xia et al., 2004) were downloaded from GenBank for analysis (Table 1). An unpublished dataset of partial 12S rRNA sequences (∼430 bp) from 19 Nephropidae (clawed lobsters) species (Tshudy et al., 2005) was also included in the analysis. These datasets were chosen because they represented different rRNA genes (12, 16, 18, 26 and 28S) from different groups of living organisms, including archaea (Aralah et al., 1996), bacteria (Rickard et al., 2002), plants (Ro et al., 1997) and animals (de Bellocq et al., 2001; Shull et al., 2001; Xia et al., 2003; Tshudy et al., 2005), with taxonomic levels ranging from class to species. The methods used in analyzing the datasets in the original published papers incorporated the common approaches of phylogenetic reconstruction, including neighbor joining (NJ), maximum parsimony (MJ) and maximum likelihood (ML), among others. The number of taxa in each dataset ranged from 19 to 49 taxa.

As a first step of our analysis, the length of each sequence was checked against the others in the same dataset and any excessive sequences from individual taxa were excluded from analysis. Our approach based on CVs was originally applied to analyze all protein sequences from complete genomes (Chu et al., 2004; Qi et al., 2004a,b) and the vectors were analogous to the peptide frequency vectors used by Stuart et al. (2002a,b). In the present study we adopted the approach in analyzing nucleotide sequences of rRNA genes. Briefly, for a sequence of rRNA gene of length L, the frequency of the appearance of oligonucleotide strings of a fixed length K was calculated. The total number of N possible types of such strings was and the total number of K-strings was . The frequency of each of the N kinds in a given DNA sequence was determined by sliding through the sequence, shifting one nucleotide position at a time. The observed frequency of a string was of a string which appeared in the given DNA sequence was calculated. The correlation between two sequences A and B was determined by taking the projection of one vector on another, and the distance between the two was defined as . After constructing a distance matrix for all sequences in a dataset, the NJ (Saitou and Nei, 1987) analysis implemented in Phylip 3.63 (Felsenstein, 1989) was used to construct the phylogenetic tree for the dataset. The details of this method were described in Qi et al. (2004a) and Chu et al. (2004) in analyzing the amino acid sequences from complete genomes of prokaryotes and chloroplast genomes, respectively. In the previous studies, in order to diminish the influence of random neutral mutations at the molecular level and to highlight the shaping role of selective evolution, such random background was subtracted from the frequencies of oligopeptide strings using a Markov model of order (K = 2) before computation of the CVs. This procedure of subtracting random background was omitted in the present study because of the limited length of the rRNA genes. Preliminary analysis on the rRNA datasets used also showed that the procedure did not further enhance the reliability of the method. The analysis was implemented using CVTree beta 1.0 which can be downloaded from http://cvtree.cbi.pku.edu.cn (Qi et al., 2004b).

To determine the length of string (K) used in the CV analysis, we followed Pevzner’s (2000) result that the best K value for a sequence of length L is . The K values used for each of the rRNA datasets to generate the distance matrices ranged from 8 to 11 (Table 1). Our preliminary analyses showed that for most datasets, the number of correctly grouped taxa reached a peak when K was above 7–8. The CV trees generated from the distance matrices were then compared with the corresponding trees constructed based on traditional methodologies with sequence

### Table 1. List of organisms (with the number of taxa in parentheses), source references, sequence information, K values used in analysis and the tree topology test results of the seven datasets studied

<table>
<thead>
<tr>
<th>Organisms</th>
<th>Reference</th>
<th>rRNA gene</th>
<th>Aligned sequence length (bp)</th>
<th>K</th>
<th>KH\textsuperscript{a}test</th>
<th>SH\textsuperscript{b}test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nephropidae (19)</td>
<td>Tshudy et al., 2006</td>
<td>12S</td>
<td>431</td>
<td>8</td>
<td>0.886</td>
<td>0.426</td>
</tr>
<tr>
<td>Arachaeae (34)</td>
<td>Arahal et al., 1996</td>
<td>16S</td>
<td>1480</td>
<td>10</td>
<td>0.310</td>
<td>0.151</td>
</tr>
<tr>
<td>Bacteria (47)</td>
<td>Rickard et al., 2002</td>
<td>16S</td>
<td>611</td>
<td>9</td>
<td>0.003\textsuperscript{a}</td>
<td>0.001\textsuperscript{a}</td>
</tr>
<tr>
<td>Adefphaga (49)</td>
<td>Shull et al., 2001</td>
<td>18S</td>
<td>2424</td>
<td>11</td>
<td>0.172</td>
<td>0.091</td>
</tr>
<tr>
<td>Tetrapoda (48)</td>
<td>Xia et al., 2003</td>
<td>18S</td>
<td>1958</td>
<td>10</td>
<td>0.009\textsuperscript{a}</td>
<td>0.006\textsuperscript{a}</td>
</tr>
<tr>
<td>Ranunculaceae (35)</td>
<td>Ro et al., 1997</td>
<td>26S</td>
<td>1063</td>
<td>10</td>
<td>0.243</td>
<td>0.119</td>
</tr>
<tr>
<td>Nematoda (20)</td>
<td>de Bellocq et al., 2001</td>
<td>28S</td>
<td>550</td>
<td>9</td>
<td>0.012\textsuperscript{a}</td>
<td>0.011\textsuperscript{a}</td>
</tr>
</tbody>
</table>

\textsuperscript{a}Kishino-Hasegawa test (Kishino and Hasegawa, 1989).

\textsuperscript{b}Shimodaira-Hasegawa test (Shimodaira and Hasegawa, 1999).

\textsuperscript{c}Significant difference in tree topologies between the CV tree and the published trees (P < 0.05).
alignment, using the Kishino–Hasegawa (KH) test (Kishino and Hasegawa, 1989) and Shimodaira–Hasegawa (SH) test (Shimodaira and Hasegawa, 1999) in PAUP 4.0 (Swofford, 2000). If the topologies of the two trees from the same dataset were significantly different, we varied the $K$ value to search for a $K$ value that could generate a CV tree that matched better with the tree constructed based on sequence alignment. For the published datasets, the trees in the corresponding papers were used for comparison. For comparison with the CV tree based on the unpublished Nephropidae dataset, a NJ tree was constructed. First, the Nephropidae sequences were aligned using the multiple-alignment program Clustal W 1.5c (Thompson et al., 1994) with adjustments made by eye. The NJ analysis with bootstrap value of 1000 was then implemented using Mega 3 (Kumar et al., 2004) based on Kimura 2-parameter distance model (Kimura, 1980).

**RESULTS**

The 12S rRNA (~430 bp) dataset including 18 Nephropidae (Arthropoda, Crustacea, Decapoda) species and Neoglyphea inopinata as outgroup (GenBank accession nos DQ298420–DQ298438) was generated by the first author of this paper and his collaborators in phylogenetic studies of this family (Tshudy et al., 2005). In both the NJ and CV trees (Fig. 1), species in the same genus always grouped together. The topologies and the relationships between taxa were also highly similar in the two trees. The only difference was in the position of *Thymopides grobovi* and *Nephropides caribaeus*. In the CV tree, they clustered as a group.

![Fig. 1. Distance trees of Nephropidae based on 12S rRNA constructed with (a) the NJ method and (b) CV analysis ($K = 8$). Numbers on branches in (a) indicate bootstrap values (1000 replicates) from NJ analysis.](http://bioinformatics.oxfordjournals.org/)

K.H.Chu et al.

1692

by guest on July 13, 2015

http://bioinformatics.oxfordjournals.org/
most closely related to the two \textit{Eunephrops} species. In the NJ tree, the former two species did not form a clade; instead there was a weak (BP < 60) bootstrap support for the grouping of \textit{N. caribaeus} with \textit{Eunephrops} spp., with \textit{T. grobovi} as the most distant taxon among the four species. Nevertheless, there was no significant difference in tree topology between the two trees based on both the KH and the SH tests ($P < 0.05$).

Arahal et al. (1996) attempted to identify 22 halophilic Archaea strains in family Halobacteriaceae collected from Dead Sea. The strains were divided into five groups based on their phenotypic features. A 16S rRNA (~1400 bp) NJ tree was then constructed on one representative strain (E1, E2, E8, E11 and E12) from each group and 27 well-known Halobacteriaceae strains, along with two outgroup taxa. The five unknown strains were assigned into three genera, \textit{Haloferax}, \textit{Haloarcula} and \textit{Halobacterium} in the NJ tree. The assignment of the five strains was identical in the tree generated from our CV analysis on the same dataset (Fig. 2). The grouping of the 27 well-known strains at the genus level was also the same in our tree and the published tree. The relationships between different genera were not well resolved in both trees, and no significant difference between the two trees was found in the KH and the SH tests.

In a study on freshwater biofilm bacteria by Rickard et al. (2002), 15 gram-negative strains were identified to the genus level based on ML analysis of partial 16S rRNA sequences (~650 bp) using data from 31 known bacterial strains from five taxonomic groups, \textit{Zymomonas}, \textit{Methyllobacterium}, \textit{Bradyrhizobium}, \textit{Rhodobacter} and \textit{Pseudomonas}. In our CV tree (Fig. 3), all bacterial strains, including the 15 unknowns, were assigned correctly into their corresponding taxonomic groups as in the ML tree. In both trees,
**Pseudomonas** group was the most distant taxon and the relationships between the other four groups were not well resolved, although the topologies were different between the two trees. While *Rhodobacter* was the sister group of a cluster consisting of *Bradyrhizobium*, *Methylobacterium* and *Zymomonas* in the ML tree, it clustered with the *Zymomonas* group only in the CV tree, with *Bradyrhizobium* and *Methylobacterium* as their sister groups. Unlike the previous two comparisons, the KH and the SH tests ($P < 0.05$) showed that the topologies of the two trees were significantly different. By varying $K$ values from 4 to 20 in our analysis, we could not generate a topology that gave better match with the ML tree.

Shull *et al.* (2001) explored the phylogenetic relationships of 36 adephagan beetles (Arthropoda: Insecta: Coleoptera) and 13 outgroup species based on full-length 18S rRNA (~2400 bp) sequences using two tree reconstruction approaches, POY (Gladstein and Wheeler, 1996, ftp.amnh.org/people/wheeler/poy) + parsimony searches and ML with 5:1 alignment weight. Broadly similar tree topologies from two approaches suggested that suborder Adephaga was a well-supported group, and monophyly of each of the two groups, Geadephaga (in terrestrial habitat) and Hydradephaga (in aquatic habitat), within Adephaga was also supported. However, monophyly of the families within the two groups was weakly supported. Moreover, family Trachypachidae which is terrestrial but possesses some features that characterize Hydradephaga, was grouped with Geadephaga. In the CV tree (Fig. 4), the clustering of monophyletic groups, including Adephaga, Geadephaga and Hydradephaga was identical to the trees of Shull *et al.* (2001).

Similarly, the monophyly of the families was not supported in our CV analysis, although the relationships between specific taxa might be different. Similar to the ML tree, Trachypachidae was placed within Geadephaga. There was no significant difference in tree topologies between our CV tree and the ML tree according to KH and SH tests.

Studies based on 18S rRNA sequences of tetrapods always supported the grouping of birds and mammals (Hedges *et al.*, 1990; Rzhetsky and Nei, 1992; Huelsenbeck *et al.*, 1996) in contrast to the grouping of birds and reptiles based on morphological, paleontological and other molecular data (Carroll, 1988; Eernisse and Kluge, 1993; Hedges, 1994). Xia *et al.* (2003) attempted to resolve this
issue by analyzing 47 tetrapod 18S rRNA sequences (~2100 bp),
with Latimeria as outgroup. By presenting a FastME (Desper and
Gascuel, 2002) tree constructed using structurally aligned 18S
rRNA sequences, the authors argued that the bird–mammal group-
ing was because of sequencing errors and misalignment of the
sequences in previous analyses. Xia et al.’s (2003) tree showed
that, other than those sequences of Hedges et al. (1990), which
included three reptiles, three amphibians and one bird (Turdus),
birds and reptiles did group together. Sequences from Hedges
et al. (1990), however, clustered together as a sister group of the
bird–reptile clade in the FastME tree. According to Xia et al.
(2003), the sequences from Hedges et al. (1990) were poor in quality
for alignment, so that those sequences failed to be assigned to the
respective amphibian, reptile or bird clades as a result of analytical
errors. In our CV tree (Fig. 5), most of the taxa could be grouped to
their corresponding amphibian, reptile, bird or mammal clades,
including Turdus that was correctly grouped to the bird clade as
the most distant taxon, rather than to the other sequences of Hedges
et al. (1990) as in the FastME tree of Xia et al. (2003). However, the
sequences of reptiles and amphibians from Hedges et al. (1990)
were grouped into a single clade, distinct from the rest. In contrast to
the FastME tree, the CV tree supported the affinity between birds
and mammals as in many previous analyses based on 18S rRNA
(Hedges et al., 1990; Rzhetsky and Nei, 1992; Huelsenbeck et al.,
1996) but not the bird–reptile relationship. When the sequences
from Hedges et al. (1990) were excluded from our analysis, the
topology of the CV tree remained the same (tree not shown). Sig-
nificant difference between the topologies of the FastME tree and
the CV tree was found based on KH and SH tests. Varying K values
from 4 to 20 in our analysis did not yield a tree topology that

Fig. 4. CV tree (K = 11) based on the 18S rRNA dataset of adephagan beetles analyzed by Shull et al. (2001).
matched better with the FastME tree. Birds and mammals always grouped together in our CV trees.

Ro et al. (1997) used partial 26S rRNA sequences (~1100 bp) of 31 Ranunculaceae (Anthophyta: Angiospermae: Ranunculales) taxa and four Berberidaceae outgroup taxa to resolve the phylogenetic relationships in Ranunculaceae at subfamily level. The traditional classification system in Ranunculaceae is based on fruit types, flower parts, the number and shape of chromosomes. The Ranunculus group (R-chromosome group) has large and long chromosomes with a base number of 8, and the Thalictrum group (T-chromosome group) has short and small chromosomes with a base number of 7 or 9. Based on NJ analysis of 26S rRNA, Ro et al. (1997) re-examined the traditional classification system and proposed four subfamilies of Ranunculaceae: (1) Hydrastidoideae, with genus Hydrastis; (2) Coptidoideae, with Coptis and Xanthorhiza; (3) Thalictroideae consisting of all T-chromosome taxa, except Hydrastis, Coptis and Xanthorhiza and (4) Ranunculoideae including all R-chromosome taxa. Hydrastis, which was placed in family Hydrastidaceae by Hoot (1991, 1995), was treated as a highly autopomorphic lineage and included within family Ranunculaceae by Ro et al. (1997). Similarly, our CV analysis also separated Hydrastis as the basal branch of this family (Fig. 6).

Monophyly of Xanthorhiza-Coptis, Trollius-Adonis, Consolida-Delphinium and Actaea-Cimicifuga-Eranthis as supported by NJ analysis was also evident in the CV tree. Yet monophyly of Ranunculus-Trautvetteria that was strongly supported by the NJ analysis

Fig. 5. CV tree ($K = 10$) based on the 18S rRNA dataset of tetrapods analyzed by Xia et al. (2003).
was not supported by the CV tree. Moreover, the position of the Consolida-Delphinium group which belongs to R-chromosome group (equivalent to Ranunculoideae) was different between the two trees. In the NJ tree, it was the sister group of a clade consisting of the other Ranunculoideae taxa and Thalictrioideae. However, in our CV tree, the Consolida-Delphinium clustered with Thalictrioideae instead of the other Ranunculoideae taxa. In any case, the phylogenetic position of this group in Ranunculaceae has always been controversial (Tamura, 1993; Jensen, 1995). Other than these discrepancies, the grouping and topology of the NJ tree and CV tree were identical. And no significant difference was found in the KH and SH tests between the NJ and CV trees.

Traditional classifications of nematodes are always problematic because reliable morphological characters are difficult to be examined. de Bellocq et al. (2001) attempted to use partial 28S rRNA sequences (~600 bp) to resolve phylogenetic relationships of 19 nematode species from Trichostrongyloidea and Strongylina groups, in the order Strongylida. In their MP tree, Trichostrongyloidea constituted a monophyletic group, and the three-superfamily classification of Trichostrongyloidea, Heligmosomoidea and Molineoidea within this group was well-supported, with Heligmosomoidea and Molineoidea as sister taxa. The phylogenetic relationships within Trichostrongyloidea could not be well resolved, with bootstrap support of only ~60. In contrast to Trichostrongyloidea, Strongylina was found to be paraphyletic in MP analysis as Triodontophorus serratus (family Strongylidae) clustered with Trichostrongyloidea (~60 bootstrap support) instead of to the other Strongylina. In the CV tree (Fig. 7a), the monophyly of Trichostrongyloidea and paraphyly of Strongylina were also evident. However, in two of the three superfamilies (Trichostrongyloidea and Molineoidea) in Trichostrongyloidea, members of the same superfamily did not cluster together, although members of Heligmosomoidea and Molineoidea were found to be closely related as in MP tree. Significantly difference in topology in the KH and the SH tests was found between the CV and MP trees. Interestingly, when the string length was lowered to four (Fig. 7b), Trichostrongyloidea became a monophyletic group and the monophyly of the three families (Haemonchinae, Cooperiidae

Fig. 6. CV tree \((K=10)\) based on the 26S rRNA dataset of Ranunculaceae analyzed by Ro et al. (1997).

![Diagram of the CV tree based on the 26S rRNA dataset of Ranunculaceae](image-url)
and Trichostrongylidae) was supported as in the MP tree. Yet the relationship among members of the other two superfamilies were identical between the CV trees of $K = 10$ and $K = 4$, but different from the MP tree. KH test showed that the CV tree of $K = 4$ was not significantly different from the MP tree, but significant difference between the two trees was found based on SH test ($P < 0.05$).

**DISCUSSION**

Qi et al. (2004a) demonstrated the applicability of CV analysis without sequence alignment in phylogenetic reconstruction of complete protein sequences from prokaryote genomes. This method has subsequently been applied, in some cases with modifications, in analyzing the chloroplast (Chu et al., 2004; Yu et al., 2005) and mitochondrial genomes (Z. G. Yu et al., unpublished data). A similar approach has previously been applied in analyzing mitochondrial genomes of vertebrates (Stuart et al., 2002a,b). While all previous analyses have been based on protein sequence analysis, using different procedures for subtraction of random background, the present study is a first attempt to apply this approach in analyzing short DNA sequences from single genes. We have used the simplest version of CVs (without subtraction of random background) in analyzing a range of rRNA datasets from different taxonomic groups of organisms available in the literature. The purpose is to test the feasibility of this approach in clustering short DNA sequences. By circumventing the sequence alignment procedure, we hope that this approach would facilitate the use of various rRNAs as molecular barcodes in species identification.

In the datasets analyzed, the groupings of rRNA sequence to taxa as revealed by our analysis, such as grouping of spiny lobsters (Nephropidae) to genera, of buttercups (Ranunculaceae) species to subfamilies, and of tetrapods to classes, are often very similar.
to those using traditional approaches based on sequence alignment. And in the Archaea and bacteria datasets, the assignment of unknown taxa to their respective taxonomic groups is identical among the two kinds of analyses. Among the seven datasets examined, the Nephropidae (12S rRNA), archaea (16S rRNA), aphagan beetle (18S rRNA) and Ranunculaceae (26S rRNA) datasets gave very similar results in terms of topology between the published trees and the trees based on analysis of CVs. This indicates that our analysis could also elucidate the phylogenetic relationship among the higher taxa, other than grouping of sequences to lower taxa. In the other three datasets, i.e. bacteria (16S rRNA), tetrapod (18S rRNA) and nematode (28S rRNA), there are significant differences in topology between the published trees and the CV trees, suggesting that the relationship among the higher taxa revealed by the two kinds of analysis are different in some cases.

In the case of tetrapods, although our CV tree is different from the Xia et al.’s (2003) FastME tree, it is very similar to those in previous studies based on 18S rRNA (Hedges et al., 1990; Zhetskay and Nei, 1992; Huelsenbeck et al., 1996) as well as based on other algorithms used by Xia et al. (2003), in which birds and mammals are grouped together. Thus, the reason why 18S rRNA always gives a topology distinct from those revealed by other datasets in tetrapods remains an issue to be explored. Yet it is interesting to note that in our tree one of the ‘poor’ sequences of Hedge et al. (1990), from the bird from Turdus, cluster with those of other birds, suggesting that our approach may be useful in analyzing ‘poor’ DNA sequences. For the tetrapod and bacteria datasets, despite the differences in tree topology, our analysis could accurately cluster the sequences to taxonomic groups (genera and classes, respectively), suggesting that the unaligned rRNA genes could serve as DNA barcodes in grouping of sequences to taxa and assigning unknown sequences to the taxa. However, the CV analysis based on the nematode dataset appears to be problematic in terms of the capability of grouping sequences together to the right taxa. Interestingly, the CV tree based on a DNA string length (K) of 4 yields a tree topology that matches better with the published tree than the CV tree of K = 9. We note that the nematode dataset is among those with a shorter sequence length. Among the datasets studied, three have sequences under 600 bp. The Nephropidae dataset has the shortest sequence length (~430 bp), and both the bacteria and nematode datasets are between 550–600 bp in length. The other four datasets had sequence length >1000 bp. Following Pevzner’s (2000) results, we have used small K values (8 or 9) in analyzing the three datasets with shorter sequences, as frequency data of long K-strings (10 or 11) may not provide enough information to reveal the relationships between the sequences. Yet only for the nematode dataset among the three datasets does our analysis fail to group the sequences to taxa. One parameter worth noting is the mean sequence divergence of the datasets, which is 14.5% for the bacteria dataset, 13.9% for the Nephropidae dataset and only 7.3% for the nematode dataset. The higher level of divergence in the former two datasets would provide more information for analysis and thus may explain why the topology of the corresponding CV trees is comparable with that in the published trees. And in the nematode dataset with the lowest divergence, a smaller DNA string length (K = 4) would enhance the resolving power of CV analysis. To elucidate this issue, the relationships between the length of DNA string used in the analysis and the sequence length and divergence in a dataset have to be explored. Further, the applicability of different algorithms for subtracting the random background in sequences, including Markov model (Chu et al., 2004; Qi et al., 2004a,b) dynamic language model (Yu et al., 2005) and discrete Fourier transform (Z. G. Yu, personal communication) in enhancing the reliability of analyzing short sequences of rRNA genes should also be investigated.

To sum up, we have demonstrated that the analysis of CVs based on unaligned rRNA sequences is a reliable clustering strategy for DNA barcoding purposes in a variety of taxonomic groups and systemic levels. While this approach was previously applied in analyzing complete genome data, the present study shows that it is also applicable in analyzing much shorter DNA sequences from a single gene, which is going to be the fundamental block in the massive barcode database. Ultimately the database is estimated to include as many as 65 billion bp (Hajibabaei et al., 2005). It has been estimated that it takes 1 h to align a dataset containing 100 sequences of 1500 bp each on a P4 1.8 GHz computer using ClustalW (Ebedes and Datta 2004). Yet in a simulation study of CV analysis, we could analyze 10 000 sequences of the same length in the same duration. Therefore, the approach would much expedite the barcoding analysis of large datasets. The approach may also be applied as a rapid method for cluster analysis of a massive dataset (>10 000 sequences) so that the subsets can be analyzed separately by alternative strategies.

It is worthy to note that the CV analysis could have other applications in DNA barcoding besides in cluster analysis. The determination of frequencies of DNA strings would enable easy identification of taxon-specific strings that can be used as taxon-specific probes in DNA chip for species identification (Summerbell et al., 2005). Moreover, the vector based on each sequence is unique and thus could serve as a taxon-specific signature, e.g. in the proposed Barcode of Life Data Systems Identification engine (Hajibabaei et al., 2005). The use of such vector signatures would reduce the size of the entire database from several hundred base pairs per taxon to ~10 digits per taxon. This taxon-specific code would be analogous to the ‘Code 39’ standard widely used in many industry and government barcode specifications. To conclude, we believe our approach without sequence alignment would much facilitate the development of various rRNA genes in barcoding of life.

ACKNOWLEDGEMENTS

The authors would like to thank two anonymous reviewers for their constructive comments which significantly improved the manuscript. The work described in this article was fully supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. CUHK4419/04M).

Conflict of Interest: none declared.

REFERENCES


