Comparative 16S rRNA Sequence Analysis

Suchart Chanama

Department of Biochemistry, Faculty of Science, Chulalongkorn University

Abstract

Ribosomal RNAs (rRNAs) involved in the production of proteins appear to be identical in function since overall three-dimensional structure of all rRNAs shows only minor variation. However, the comparison of variable regions of rRNAs obtained from various microbial origins provides highly significant molecular information in sense of genetic relatedness or phylogenetic (evolutionary) relationships. Comparative study of nucleotide sequences of 16S rRNA has been chosen because of their appropriate length and sufficient sequence information. This review describes methods for studying genetic relatedness of microorganisms using comparative 16S rRNA sequence analysis, its application in bacterial systematic identification particularly for medically important Clostridia as well as prospective research of 16S rRNA.

Key words: 16S rRNA, phylogenetic relationships, sequence analysis
INTRODUCTION

A prime objective of bacterial systematics certainly must be the establishment of a classification that span over the prokaryote kingdom. However, classification by traditional or classical methods have been difficult and ambiguous especially beyond the level of genera, for example the technique of gram's staining that is used for classification and identification of bacteria containing cell walls, in turn mycoplasmas are cell wall-less bacteria cannot be defined by this technique (in fact, they are degenerate clostridia). Another problem is that use of morphological and physiological characteristic judgment to classify very closely related species in the same genus, such as Clostridia, in which biochemical patterns are variable, results in controversial identifying these bacteria (Fox et al., 1977; Johnson et al., 1975) Therefore, the phenotypic, physiological, and biochemical examinations are not enough for characterization. Recently, it has been proposed that this problem might be solved by using molecular-based approach, i.e. analysis of sequence information of nucleic acids and/or proteins (Fox et al., 1977; Fitch et al., 1967; Fox et al., 1980; Woese, 1987; Lake, 1988). This approach is, now, more powerful to measure at level of evolutionary relationships. Genotypic or sequence information is superior in two main ways to phenotypic information. First, sequence information is more readily, reliably, and precisely interpreted. Second, it is innately more informative of evolutionary relationships than phenotypic information is. The sequence information is also one-dimensional, compared to three-dimensional phenotypic pattern, which can be measured in simple ways. Additionally, it is an objective judgment as number (numeric value) for mathematical analysis. As results, molecular approach has been introduced to improve and resolved the way of classification. However, DNA and amino acid sequence information are not enough powerful to reveal higher bacterial taxa. Previously, information of small subunit ribosomal RNA sequences has been proven to be more useful to solve the problem (Stackebrandt et al., 1981; Lane et al., 1985; Stackebrandt et al., 1987; Ash et al., 1991; Ash et al., 1991a; Weisburg et al., 1991; Reysenbach et al., 1992; Govindaswami et al., 1993). Since it contains both very highly conserved and variable regions that little change over time which causes being smooth of evolution. It is found in all organisms including prokaryote and eukaryote. Because of its high conservatives and its smoothly changes in time, it is the most useful to solve the resolution of very related organisms, and it acts as "molecular chronometer" (Fox et al., 1977; Johnson et al., 1975; Fox et al., 1980; Woese, 1987).
Why ribosomal RNAs (rRNAs) are the ultimate molecular chronometers? At first, rRNAs show a high degree of functionally constancy, which assures relatively good clocklike behavior. Second, they occur in all organisms (ubiquitous and being clearly of very ancient origin) and different positions in their sequences change at very different rates, allowing most phylogenetic relationships (including the most distant) to be measured, which makes their range all-encompassing. Third, primary structure of these rRNA molecules are sufficiently constrained that on the whole they have not changed rapidly in time. Fourth, their size are large and they consist of many domains; there are 50 helical stalks in the 16S rRNA secondary structure and roughly twice that number in the 23S rRNA, which make them accurate chronometers on two counts. Fifth, they contain regions of both extreme conservation and hypervariability, thus both distant and close relationships can be examined. Finally, they can be sequenced directly (and rapidly) by means of enzyme reverse transcriptase.

Currently, the appropriate subunit size of ribosomal RNAs that has been chosen to be analyzed widely is 16S rRNA which is as a result of first, its length is of about 1,600 nucleotides which provides a more reliable classification than the small 5S rRNA (~120 nucleotides), and second it is experimentally more manageable than the larger 23S rRNA (~3,300 nucleotides).

HOW THE 16S rRNA IS USED TO STUDY GENETIC RELATEDNESS

To extract the information of 16S rRNA for deducing phylogenetic and evolutionary relationships among bacteria, many approaches have been developed for receiving partial or full length of the rRNA sequences as a resource of information, then analyzing these sequences by alignment and comparison of individual data with other ones, with sophisticated algorithms. Based on appropriate alignment and algorithm used for analysis, the result of homology and percentage similarity can be calculated and dendrogram (phylogenetic tree) can be produced finally. The detail of each techniques are described as follows:

A. Oligonucleotide Cataloging Method

A partial sequence characterization, in term of a comparative cataloging approach, has been used for many years since it was not feasible to determine complete rRNA sequences (Fitch et al., 1967; Fox et al., 1977; Fox et al., 1980; Stackebrandt et al., 1981; Sobieski et al., 1984; Stackebrandt et al., 1987; Woese, 1987). Cellular, total rRNAs of each bacterium are obtained by general hot acid phenol extraction and subsequently precipitated the rRNAs with
high salt concentration of NaCl (2M). The 16S rRNA is separated by polyacrylamide gel electrophoresis with final purification by passing over a Whatman CF-11 cellulose column. The second step is to construction of oligonucleotide catalog of each individual 16S rRNA from individual bacterium. Each purified 16S rRNA is digested into short oligonucleotides, of lengths up to 20 or more bases, with ribonuclease T1 (which cleaves specially at G residues) and the resulting oligonucleotides (or sequence fragments) are collected and given as an "oligonucleotide catalog". The oligonucleotides are then separated and sequenced by two dimensional paper electrophoresis technique. Briefly, the oligonucleotide digests are subject to electrophoresis on cellulose acetate at pH 3.5 (presence of 7M urea), subsequently transfer of the oligonucleotide pattern to diethylaminoethyl-cellulose paper, then orthogonal electrophoresis. The resulting oligonucleotide fingerprint comprises of particular pattern is further sequenced by digestion with various secondary and tertiary enzymes. Consequently, the sequence of each oligonucleotide is deduced therefrom. Next, the important step, data from all oligonucleotide catalogs are analyzed and expressed in terms of "percentage sequence similarity" and "binary association coefficients" (SAB value) by comparisons of any pair of catalogs. Finally, dendrograms (phylogenetic tree that reveals the genetic relatedness among bacteria) are constructed using any of several clustering algorithms. There are at present three main procedures used for this purpose, distance matrix methods, maximum parsimony methods, and cluster analysis. In essence, basic to all of these methods is some mathematical criterion that allows one particular tree to be selected from the many possible alternatives for a given set of orthologous macromolecular sequences. Choosing the phylogenetic tree that minimizes the difference between the tree distances (phyletic distances) and the observed sequence differences is a such example of these methods. In addition, the higher order structures of the 16S rRNA sequence comparisons (by multiple alignment the signature positions of 16S rRNAs) also help this classification.

B. Direct 16S rRNA Sequencing Technique

To rapidly obtain sequence data, direct 16S rRNA sequencing technique has still been used well, though it requires relatively large amounts of cellular RNAs and is prone to errors because only one strand is sequenced and problem of secondary structure of RNAs. However, because of abundant ribosomal RNAs (up to 50% of total cellular RNA) in the cell, the idea of rRNA as a template for direct sequencing was possible to perform and was
interesting many people in the past few years (Lane et al., 1985; Ash et al., 1991; Ash et al., 1991a; Reysenbach et al., 1992). This method is carried out by using enzyme avian myeloblastosis virus reverse transcriptase and based on the knowledge of conserved sequence elements among widely divergent species which are distributed along the length of 16S rRNA gene as universal primers, and dideoxynucleotide chain termination technique to sequence nearly entire this gene directly from extract RNA (cellular RNA). The sets of sequence data from individual bacteria generated by this mean serve as a oligonucleotide catalog, and analyses of these data are subsequently evaluated by the same algorithm(s) as mentioned earlier.

C. Polymerase Chain Reaction (PCR) and Direct Sequencing of 16S Ribosomal DNA

The technique of PCR is a very powerful tool, extensively used in many areas including molecular biology. This strategy offers several advantages over direct sequencing of 16S rRNA (Weisburg et al., 1991; Govindaswami et al., 1993). First it utilizes relatively crude preparations of total DNA such as from minipreparations, that is usually easier to prepare from the cell than is RNA. Second, only small amounts of DNA are required. Third, both strands of the gene can be sequenced which reduces errors, and can obtain the sequence information through the complex secondary structures found in ribosomal RNAs. Basically, this technique takes an advantage of conserved elements within ribosomal RNA sequence, which are nested the 16S rRNA coding region to initiate DNA synthesis in dideoxynucleotide sequencing reactions. By this mean full length of 16S rRNA sequence can be determined for alignment and comparisons. The evolutionary distance and percentage similarity are calculated as similar to previous methods.

CLOSTRIDIA: AN APPLICATION OF COMPARATIVE 16S rRNA SEQUENCE ANALYSIS

The genus Clostridium is not a natural taxonomic classification. It is a heterologous assembly of gram positive, catalase and oxidase-negative, strictly fermentive anaerobic, rod-shaped endospore-forming bacteria although a number have lost one or more of these characteristics. The heterogeneity of the group can be defined by range of %G+C content, which varies from 24 to 55 mol% (Johnson et al., 1975). Since no suitable techniques was available to precisely and accurately classify this genus deeply to exact species and/or strains until advent of this comparative analysis. In consequence, it is likely to re-classify most uncertain members of this genus by this technique to obtain the right taxonomic grouping.
However, there are a few papers, reported re-establishment or re-grouping some *Clostridium* spp. (Zhao *et al.*, 1990; Kane, 1991; Canard *et al.*, 1992), one of them attempted to closely relate the members of *Clostridium* into the same group by this comparative sequence analysis. The reason here are few reports in this area might be due to the fastidious behavior of anaerobic bacteria placed into this genus, which make them impossible or hard to grow in artificial culture media, thus no available pure culture to be examined and studied. This drawback was also reported by Zhao and colleagues (1990). Another reason is the difficulty of justification and/or judgment of very closely related species which exhibit very similar phenotypic, physiological and biochemical characteristics, and finally the hardship of understanding the complicated algorithms that are used mainly in phylogenetic study.

Getting started with the paper of Zhao *et al.* (1990), they attempted to resystematize the former *Clostridium bryantii*, which is a very fastidious strictly H$_2$-reducing, interspecies proton transfer-dependent, 4-carbon to 11-carbon saturated fatty acid-b-oxidizing anaerobe. It requires cocultivation with H$_2$- and formate-utilizing bacteria, *Methanospirillum hungatei*. Because its complicated metabolic process is obligately coupled to a utilization of electron acceptor, the pure culture was possibly unavailable. As a result, it is the first obstacle limiting the informative detail for further analyses and examinations. Fortunately, a recent report of successful growing pure culture of another anaerobic bacteria, *Syntrophomonas wolfei* (fatty acid-b-oxidizing syntroph) by passage adaptation in crotonate (fatty acid) medium, led them to achieve isolation of this organism. Based on reconstitution experiment, the pure crotonate-grown culture of *C. bryantii* was used to reconstitute the caproate (fatty acid which has been first defined to be used by isolated *C. bryantii*) medium containing coculture of *M. hungatei* in order to retestify the actual property of *C. bryantii* (if *C. bryantii* is really utilizing the caproate as a source of energy). Their success of isolating pure culture let *C. bryantii* be further studied and examined in detail including at molecular level. Although its DNA composition analysis (37.6 mol% G+C) and some biochemical properties shows that *C. bryantii* could be placed in the genus *Clostridium*, the physiological characteristic exhibits its behavior of syntrophic caproate degradability which is a character of syntrophic bacteria. Moreover, use of 16S rRNA sequence analysis, following the procedure of direct 16S rRNA sequencing technique, obviously revealed that *C. bryantii* is much more closely related to *Syntrophomonas wolfei* than any *Clostridium* species, with relatively lowest evolutionary distance value of 6.1 while range of these values between *C. bryantii* and any *Clostridium* spp. is 18.1-20.7. Therefore, as a consequence of
combination of the sequence homology information and its behavior as syntrophic fatty acid degrader, they recommended reclassify this bacteria to a new genus of Syntrophospora bryantii. This research is only the first case report successfully grouping without equivocal result even if they did partial sequence comparisons of the 16S rRNA, it might be that the sequence regions chosen for comparison analysis sit on the signature positions of its evolution, i.e. the sequence region(s) or domain(s) that is the representative of evolutionary changes. In addition, if full lengths of 16S rRNA sequences of all Clostridium spp. and related bacteria were analyzed, it would provide more information which might be useful to resolve the genetic related trees, and it was probably open the new world of taxon. This result also infer that if classification encounters the problem, the other suitable and more powerful approach, that can offer higher resolution of differentiation at the vague branch point of dendrogram, and this criterion should be taken into your account. The weak point and limitation of this paper is that they did not analyze and compare secondary structure of the ribosomal RNAs that is very essential to retrospect the mutual origin of the closely related C. bryantii and S. wolfei to support their result.

The second is one of interesting papers, which reported a new Clostridium species isolated from gut content of the African soil-feeding termite, Cubitermes speciosus, which shares many features among Clostridia (Kane, 1991). It is very interesting since most acetogenic organisms that are placed in the genus Clostridium are of significant importance in biotechnological industry, especially solventogenesis. Hence, discovery of new acetogen(s) tends to convince industrial microbiologists to study its detail of metabolic processes which is potentially useful for industrial improvement. Kane (1991) reported the new acetogenic Clostridium mayombei, which shares similarity of some characteristic features to that of genus Clostridium. However, this strain exhibited the difference in H₂-oxidizing/CO₂-reducing fermentation property from any normal H₂/CO₂ acetogenic Clostridium when fermenting some substrate, e.g. succinate, which resulted in harder to place it in any group of Clostridium. This bacteria showed that its 16S rRNA sequence similarity was closely relative to the Clostridium lituseburense with evolutionary distance of 5.1, but is far from other Clostridial members (evolutionary distance value: ≥ 11.8 units) and other H₂/CO₂ acetogens from wood-feeding termites. Even though this new strain is very closely relative to one member of the Clostridia, substantial difference in substrate utilization was found. Consequently, based on the typical characteristics (morphology, gram positive staining, catalase and oxidase negative, oval
endospore-forming features), DNA composition (25.6 mol% G+C content), and comparative 16S rRNA sequence analysis (5.1 evolutionary distance unit and ~95.0% sequence similarity value to *Clostridium lituseburense*), and its physiological and biochemical differences from almost all of members of homoacetogenic Clostridia, they suggested that it could be placed in an independent group or new species, *Clostridium Mayombei*. This paper was carried out in a similar way to the previous report, and they lack characterization of higher-order secondary structure of the 16S rRNA sequences. In my view point, the investigators have not realized carefully about the advantage and disadvantage or weak point of each technique they used. Therefore, they could not extract important information from the sequence data that might be due to the technique(s) cannot reach in detail or cannot be carried out successfully, for example direct sequencing from 16S rRNA (as a template) often faces the problem of higher-order structure interference (e.g. secondary structure). However, to get the best result entire sequence of this rRNA should be obtained by any method that is reliable in terms of accuracy, precision, and reproducibility; such as PCR for amplifying a short DNA sequence. If taking a look the two papers above carefully, one will see the quite differences between *Clostridium bryantii* and *Clostridium mayombei*. The former is fastidious, few substrate-utilizing, and H₂-reducing anaerobe whereas the latter is nonfastidious, many substrate-utilizing, and H₂-oxidizing anaerobe. As a result, use of general characteristics including morphology, gram's staining, biochemical reactions as criteria to examine and classify these organisms, cannot justify and give the final result since those methods cannot reach and/or approach to find detail of the organisms. The molecular-based method (16S rRNA) that mentioned in the earlier section, reaches this goal as evidences reported by authors of the papers.

The last one is concerned to medical importance, well-known *Clostridium perfringens*. Canard *et al.* (1992) had applied the technique based on 16S rRNA analysis to make up relationships of pathogenic *Clostridium perfringens* among other Clostridia via sequence alignment of about 370 rRNA sequences on the basic of conserved domains and secondary structure. With their scrutiny of aligned sequences, they strikingly observed a particular loop deletion (which acts as molecular signature for most *Clostridium*) at the positions equivalent to positions 455-479 of the *E. coli* 16S rRNA sequence (interestingly, this region is not included in the phylogenetic analysis) while none of previously published sequences for bacteria belonging to the low %G+C content cluster had such a deletion. It is likely that this deletion is found particularity in the *Clostridium* cluster. In conjunction with evolutionary distance value
calculated for each sequence pair among gram positive bacteria species, they could disclose the genetic relationships between Clostridium perfringens and Clostridium pasteurianum with relatively lowest distance value of 5.9. This is the first paper that published the phylogenetic classification of human pathogenic Clostridium based on the 16S rRNA analysis and the results are in agreement with those of previous works (based on 23S rRNA reassociation method). This paper also shows that observation of the sequences carefully is crucial for making an appropriate alignment that can exclude some less genetic related bacteria from the main cluster in order for the resolution of overall algorithms. Further investigation of this research might attempt to characterize and examine in detail of Clostridium pasteurianum which is closely relative to pathogenic C. perfringens, whether it is a potent pathogen!

PROSPECTIVE RESEARCH AND DISCUSSION

Since the most informative data for doing taxonomic grouping relies on the quantity of sequences accumulated in the database, in the next few years there is a need to persuade researchers to conduct 16S rRNA sequencing by direct method or using PCR technique. However, inclination to use PCR (with direct sequencing the products) is much higher than using direct ribosomal RNA sequencing because the length of 16S rRNA coding gene is about 1,500 nucleotides which is not prone to an error of the Taq DNA polymerase, and DNA sequencing is now straightforward and convenient to perform automatically with commercially available DNA sequencer.

In the past several years, there are a few researches conducting reclassify members of the genus Clostridium, although this genus Clostridium contains lots of members justified in many clusters (Johnson et al., 1975; Woese, 1987), including typical Clostridia cluster and other clusters which are rarely less related to the typical Clostridia (e.g. unnamed and uncertain anaerobes). So, this might be the suitable time to initiate research of reclassification of ambiguous species and strains in the Clostridium genus due to the advanced knowledge of comparative 16S rRNA sequence analysis. Furthermore, it inclines to find genetic relationships between the genus Clostridium and the bacteria outside this genus, but the same phylum, in order to construct dendogram to reveal evolutionary progress of the Clostridia and their closely related species over time, which may be important to predict genetic diversity of the Clostridium sp., especially pathogens (because if the genetic of pathogen is diverted by somewhat, e.g. pollution, that results in changes or some modifying the morphology and some metabolic
pathways without loss of virulence which causes misdiagnosis, the basic knowledge of classification by this technology can be applied to this area because of constancy of function and very slow change of the ribosomal RNA sequences).

Extensive sequencing of the ribosomal RNAs has been particularly informative both partial and full sequences of the 16S rRNA. Sequence divergence among these individual rRNAs has defined the outline of a natural classification of microorganisms. These data can also be useful to design hybridization probes for determinative studies in several areas of microbiology (Stahl et al., 1988; Amann et al., 1995; Hugenholtz et al., 1996). The 16S rRNA has been the common target for determinative hybridization probes. By using selected regions within the larger 16S rRNA molecules as hybridization targets for synthetic oligonucleotide probes, probe specificity can be freely adjusted. Microbial species or subspecies can be distinguished by oligonucleotides complementary to the most variable regions of the molecule. In contrast, by targeting regions of increasing conservation, probes can be made to encompass specific genera or higher taxons. Some regions of the rRNAs have remained essentially unchanged in all sequenced species; these can be used as targets for universal probes of organisms. This idea will be a very useful and valuable for research to develop the techniques for convenient identification, detection, and rapid monitoring microorganisms in many fields of application.

REFERENCES