Identification of unknown bacteria by 16s rRNA sequencing

Murugadas Vaiyapuri
Microbiology Fermentation and Biotechnology Division
ICAR- Central Institute of Fisheries Technology, Cochin
murugadascift81@gmail.com; Murugadas.V@icar.gov.in

What is 16s rDNA?

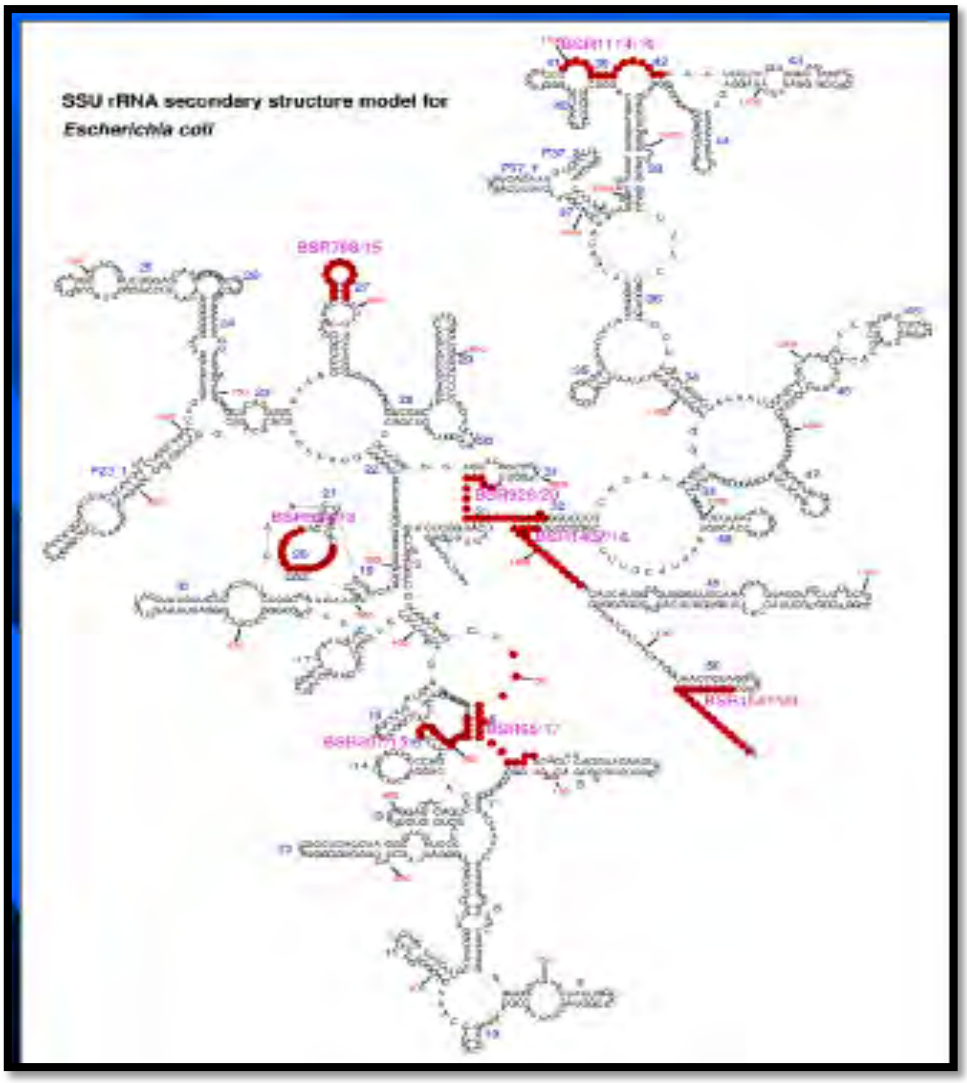
rRNA gene is the most conserved and used to determine taxonomy, phylogeny (evolutionary relationships). It is also used to infer relationships between organism that span the diversity of known life look. These genes are conserved through the billions of years of evolutionary divergence.

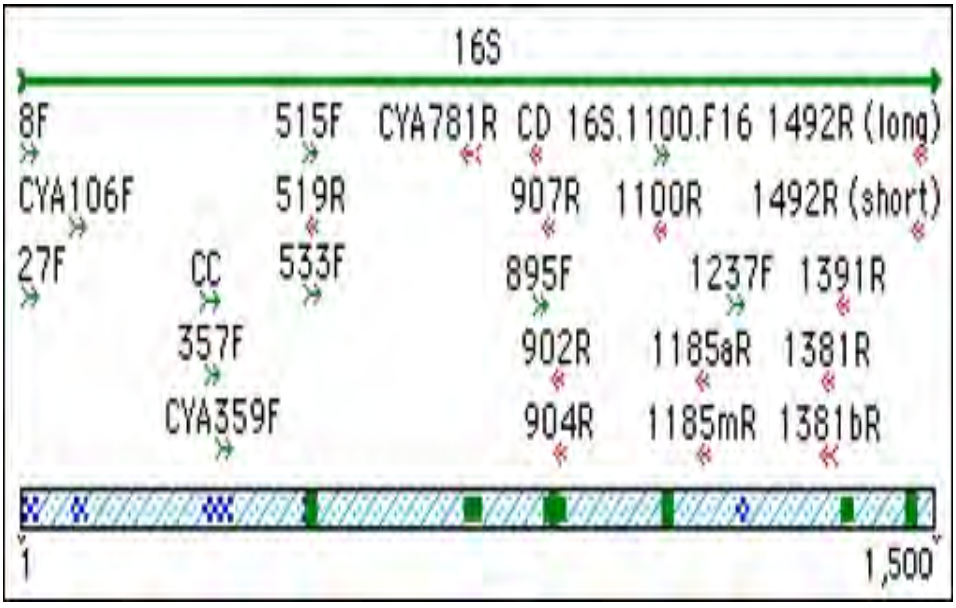
What is Microbial systematics?

Novel species is recognized using the polyphasic approach - multidimensional aspects of organisms (phenotypic, genotypic and chemotaxonomic traits). Biochemical methods of identification of bacteria are called as phenotypic methods, 16s rRNA sequencing and DNA-DNA hybridization are called as genotypic methods and other FAME analysis are chemotaxonomic methods. Phylogenetic analysis based on 16S rRNA gene sequences and determination of similarity between sequences - first step in identifying novel organisms - most widely used methodology in the world. DNA–DNA hybridization (DDH), which measures indirectly the degree of genetic similarity between two genomes, has been the 'gold standard'

What is 16s rRNA sequencing analysis?

One or more copies of the operon dispersed in the genome (mostly 3, *E. coli* 7). Ribosomal RNAs in Prokaryotes: 5S 120 Large subunit of ribosome; 16S 1500 Small subunit of ribosome; 23S 2900 Large subunit of ribosome. The 16s rDNA sequence has hypervariable regions, where sequences have diverged over evolutionary time. Strongly conserved regions often flank these hypervariable regions. Primers are designed to bind to conserved regions and amplify variable regions. Numbered primers are named for the approximate position on the *E. coli* 16S rRNA molecule. More details can be sought from National Center for Biotechnology Information (www.ncbi.nlm.nih.gov) and the Ribosomal Database Project (http://rdp.cme.msu.edu/). Minimum: 500 to 525 bp sequenced; ideal: 1,300 to 1,500 bp sequenced 1% position ambiguities. minimum: 99% sequence similarity; ideal: 99.5% sequence similarity. Sequence match is to type strain or reference strain of species that has undergone DNA-relatedness studies. For matches with distance scores 0.5% to the next closest species, other properties, including phenotype, should be considered in final species identification.


SSU rRNA secondary structure model for
Escherichia coli


16S
8F
515F
CYA781R
CD
16S.1100.F16
1492R (long)
CYA106F
519R
907R
1100R
1492R (short)
27F
CC
533F
895F
1237F
1391R
357F
902R
1185aR
1381R
CYA359F
904R
1185mR
1381bR
1
1,500

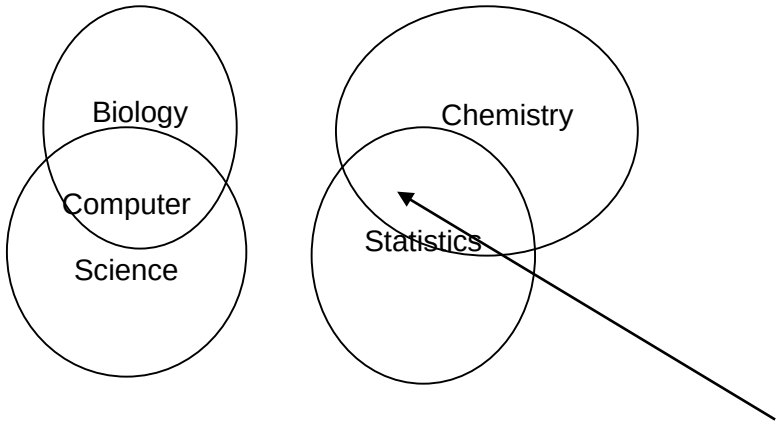
Primer		
Name	Sequence[a]	Amplified hypervariable region
V3F	5′ CCAgACTCCTACGGGAGGCAG 3′ (334–354)	V3 (334–537)[b]
V3R	5′ CGTATTACCGCGGCTGCTG 3′ (519–537)	
V6F	5′ TCGAtGCAACGCGAAGAA 3′ (961–78)	V6 (986–1043)
V6R	5′ ACATtTCACaACACGAGCTGACGA 3′ (1062–85)	
Molecular beacon probe[c]		
Name	Sequence	Target region
SEP-V6 probe	TxR-5′ tgcgcCTAGAGGGGTCAGAGGATgcgca 3′-BHQ2	1005–1022*

Other genes for identification or differentiation of bacteria?

- 23s
- 16s-23s ITS
- *rpoB*
- *gyrB*
- *Hsp*
- *recB*

What is bioinformatics?

Bioinformatics is a new science that uses computational approaches to answer biological questions. Bioinformatics is a new scientific discipline created from the interaction of biology and computer. Biological questions raised from the researchers will be investigated with the large & complex data sets available in public as well as generated by the own laboratory in private to arrive at a valid biological conclusion.



The National Center for Biotechnology Information (NCBI) defines bioinformatics as: "Bioinformatics is the field of science in which biology, computer science, and information technology merge into a single discipline"

Broad Areas in Bioinformatics

- Genomics

- Proteomics
- others

Some of the bioinformatics applicable are

Similarity search

- Sequence comparison: Alignment, multiple alignment, retrieval
- Sequence's analysis: Signal peptide, transmembrane domain,
- Protein folding: secondary structure from sequence
- Sequence evolution: phylogenetic trees

Important terms in Bioinformatics

Fasta sequences

The FASTA format is used in a variety of molecular biology software suites. In its simplest incarnation (as shown above) the ''greater than'' character (>) designates the beginning of a new file. An identifier (L04459 in the first of the preceding examples) is followed by the DNA sequence in lowercase or uppercase letters, usually with 60 characters per line. Users and databases can then, if they wish, add a certain degree of complexity to this format. For example, without breaking any of the rules just outlined, one could add more information to the FASTA definition line, making the simple format a little more informative, as follows:

>gi|171361|gb|L04459|YSCCYS3A Saccharomyces cerevisiae cystathionine gamma-lyase (CYS3) gene, complete cds.

GCAGCGCACGACAGCTGTGCTATCCCGGCGAGCCCGTGGCAGAGGACCTCGCTT
GCGAAAGCATCGAGTACCGCTACAGAGCCAACCCGGTGGACAAACTCGAAGTCA
TTGTGGACCGAATGAGGCTCAATAACGAGATTAGCG

Similarly, the protein record in fasta as follows

>P31373

MTLQESDKFATKAIHAGEHVDVHGSVIEPISLSTTFKQSSPANPIGTYEYSRSQNPNRE
NLERAVAALENAQYGLAFSSGSATTATILQSLPQGSHAVSIGDVYGGTHRYFTKVAN
AHGVETSFTNDLLNDLPQLIKENTKLVW

Majority of the procedure analysing either DNA or Protein sequences involves the use of fasta format

Practical on Blast analysis and identification of unknown bacteria from 16s rRNA gene sequence data

1. Check for the quality of sequence data with chromatogram file and pdf
2. Check the quality value of each sequence base call in the chromatogram file
3. Trim the sequence according to the sequence data quality value more than 20
4. Blast analysis of raw and trim sequence data in NCBI_Blast_nucleotide.
5. Perform merging with emboss merger after reverse complementing the reverse sequence data
6. Avoid the low-quality sequences in the analysis.

References

1. Andreas D. Baxevanis and B. F. Francis Ouellette. BIOINFORMATICS-A Practical Guide to the Analysis of Genes and Proteins. SECOND EDITION. Wileys Inerscience, A JOHN WILEY & SONS, INC., PUBLICATION. 2001.

2. Des Higgins and Willie Taylor. Bioinformatics-Sequences, structure, and databanks- A practical approach. Oxford University Press. 2000.
3. Andrzej Polanski and Marek Kimmel. Bioinformatics. Springer-Verlag Berlin Heidelberg 2007