Exercise 13

Phylogeny of Prokaryotes based on 16S rRNA genes

Advisor

Thomi Horath, horath@botinst.unizh.ch, 01 634 82 27

Reading

Chapters in BBOM 9th: 12.4 - 12.7
Chapters in BBOM 10th: 11.4-11.8
BBOM: Madigan M.T., J.M. Martinko and J. Parker: "Brock - Biology of Microorganisms", 9th edition, 1999. 10th edition, 2003. Prentice Hall.

Objectives

- Get to know some useful sequence analysis databases available on the internet.
- Get to know a computer-based resource for sequence analysis.
- Why did the 16S rRNA molecule become so famous?
- Why sequence analyses conducted on different databases yield different results.

Background

Originally, the taxonomic **classification of prokaryotes** depended exclusively on **phenotypic characteristics** such as

- **shape**: coccus, rod, spirillum, vibrio, etc. (BBOM 9th, 3.4; BBOM 10th, 4.4),
- **motility**: not motile, movement by gliding, with flagellum or flagella (monotrich, polytrich, peritrich), etc. (BBOM 9th, 3.11, BBOM 10th, 4.10-4.11)
- **behavior**: chemotactic, phototactic, magnetotactic (BBOM 9th, 3.12; BBOM 10th, 4.12)
- **membrane structure**: e.g. ester-lipids vs. ether-lipids (BBOM 9th, 3.5; BBOM 10th, 4.5),
- **cell inclusions and surface structures**: slime layers, capsules, glycogen, sulfur, magnetosomes, spores etc. (BBOM 9th, 3.13-3.15; BBOM 10th, 4.13-4.15)
- **metabolism**: phototrophic, chemotrophic, lithotrophic, organotrophic, autotrophic, heterotrophic (BBOM 9th, 4.1; BBOM 10th, not available)
- **resistance** to antibiotics (BBOM 9th, 18.12; BBOM 10th, 20.12)
- **cell wall**: Gram-positive or Gram-negative, LPS (BBOM 9th, 3.7, 3.8; BBOM 10th, 4.8-4.9)
- **pathogenicity, virulence** etc. (BBOM 9th, 1.7, 1.8; BBOM 10th, 1.4, 1.5)
- **and other characters:** Pigments, temperature tolerance, ecotype, etc.

Some of these properties turned out to be good distinguishing characteristics (e.g., the Gram stain), while others were not (e.g., cell shape).

A few years ago, **genotypic classification** based on nucleotide sequence comparison of **16S ribosomal RiboNucleicAcid** (**16S rRNA**) genes became available as an additional taxonomic tool (BBOM 9th chapters 12.4 & 12.5; BBOM 10th, 11.4, 11.5). 16S rRNA, along with the 23S rRNA, has properties which predestine it as a **universal phylogenetic marker**. All living organisms, prokaryotes as well as eukaryotes, contain the small (16S or 18S) and the large (23S or 28S) subunit ribosomal RNA and they always have the same function (BBOM Fig. 12.7 BBOM 10th, Fig. 11.8). Ribosomal RNA must have been present since very early in the development of life, because it is essential for protein synthesis. Any mutation in the 16S rRNA gene directly affects the proper functioning of the ribosome and leads to the elimination of less efficient organisms through selection. One may assume, therefore, that the 16S rRNA genes contain a large number of **highly conserved sequence patterns**. There

are a number of sequence differences which did not impair on the functioning of the ribosome, however, and which were maintained over evolutionary times. These can be used to distinguish **phylogenetically different organisms**.
There are regions on the 16S rRNA which are quite conserved and others which are variable. Comparing the differences in the **base sequence of 16S rRNA genes** is, therefore, a good tool to detect evolutionary changes and phyolgenetic relatedness of organisms. The question remains, however, how many point mutations might have occurred at one base position up to now, from A to G to U and back to A for instance. This uncertainty and the fact that we do not know when the particular mutations took place make it difficult to use the 16S rRNA as an exact **evolutionary clock**.

One branch of **bio-informatics** is studying the relatedness between organisms based on sequence comparison. The molecular sequence data obtained worldwide from sequencing projects are collected in **databases** and made publicly available. With this information, one can test novelty and possibly function of new sequences, search for homologous patterns and regulatory domains and create hypothetical phylogenetic relationships called evolutionary trees. The information which is encoded in sequences can be analyzed by comparing against existing sequences whose functions are already known. **The computer uses algorithms to find similarities between the query sequence and every sequence in the database and scores them according to the degree of relative similarity**. Different databases might lead to differences in scoring depending on the algorithm and the dataset used. It is reasonable, therefore, to conduct searches on at least two different database libraries in order to obtain the best possible homologous sequences.

As a course exercise, we will use the database provided by the National Center for Biotechnology Information (NCBI) and apply the **BLAST algorithm** to the nucleotide sequences given below (BLAST = Basic Local Alignment Search Tool). The BLAST algorithm searches for patches of similarity, and it is based on ungapped sequence alignments, i.e. the alignments created by BLAST do not allow for gaps, but BLAST does allow multiple hits on the same sequence. With this type of statistical model one increases search speed but one reduces sensitivity, i.e. BLAST might miss certain matches.

The basic concept underlying all **comparing and tree building methods** is this: Take two sequences, make them the same length and divide the number of identical nucleotides by the number of all nucleotides in one sequence (and gaps if there are any). What you get is the basic similarity percentage between two sequences (BBOM 9th Fig. 12.9; BBOM 10th, 11.10). Presently, similarities of less than 95% define different genera, less than 97% different species.

Literature

- Stackebrandt, Erko and Michael Goodfellow (eds.) 1991. "Nucleic Acid Techniques in Bacterial Systematics", John Wiley & Sons; Chichester, New York, Brisbane, Toronto, Singapore. ISBN: 0-471-92906-9
- Peruski Jr., Leonard F. and Anne Harwood Peruski. 1997. The Internet and the New Biology - Tools for Genomic and Molecular Research . American Society for Microbiology, Washington DC. ISBN 1-55581-119-1

www. Links

- U.S. National Library of Medicine and National Institutes of Health; **National Center for Biotechnology Information (NCBI)**: http://www.ncbi.nlm.nih.gov/

- The **Ribosomal Database Project (RDP):** http://rdp.cme.msu.edu/html/index.html

- The **European Molecular Biology Laboratory**: http://www.embl-heidelberg.de/
- EMBL's **European Bioinformatics Institute (EBI)**: http://www.ebi.ac.uk/index.html
- The **Oligonucleotide Probe Base:** http://www.microbial-ecology.de/probebase/index2.html

Exercises

Below you will find two 16S rDNA nucleotide sequences. Try to find out the following, applying the NCBI internet resources:

1. What are the sequences available in BLAST that share the highest relative level of similarity with the sequences given?
2. To which organisms do sequences 1 and 2 belong to?
3. What is the minimal length of nucleotides to be typed into the BLAST window to still get the correct nearest organism?
4. Is this length always the same or does it depend on the fragment location which you choose from the entire sequence?
5. Take a small fragment of your favorite sequence and exchange one or more nucleotides in it. Does this new sequence still have the same nearest relatives?
6. On the Ribosomal Database Project it is possible to create sequence based trees. Can we fit the inhabitant of lake Cadagno into its adequate group?

Genomic sequences to be analyzed with BLAST:

1. **A famous inhabitant of lake Cadagno:**

CGTGGCGGTATGCTTAACACATGCAAGTCGAACGTCAAAGGTCTTCGGATTGAGTAG
CGTGGCGGACGGGTGAGTAAAGCGTGGGAATCTGCCTTGCAGTGGGGGATAACCCG
GGGAAACTCGGGCTAATACCGCATACGCCCTACGGGGGAAAGGGGGCTTTGGCTCT
CGTTGCAAGATGAGCCCACGTCCGATTAGCTAGTTGGTAGGGTAAAGGCCTACCAAG
GCGACGATSGGTCGCTGGTCTGAGAGGATGACCAGCCACACTGGGACTGAGACACG
GCCCAGACTCCTACGGGAGGCAGCAGTGGGAATATTGGACAATGGGGGAAACCCTG
ATCCAGCAATACCGCGTGTGTGAAGAAGGCCTGCGGGTTGTAAAGCACTTTCAGTGG
GAAAGAAAACCTGGTGGTTAATACCCATCGGCTTTGACGTTACTCACAAAAGAAGCAC
CGGCTAACTCCGTGCCAGCAGCCGCGGTAATACGGAGGGTGCAAGCGTTAATCGGA
ATTACTGGGCGTAAAGCGCACGTAGGCGGCGCCGTCAGTCCGATGTGAAAGCCCTG
GGCTTAACCTGGGAACTGCATTGGATACTGCGGCGCTAGAATGTGAAAGAGGGGAGT
GGAATTCCAGGTGTAGCGGTGAAATGCGTAGAGATCTGGAGGAACACCAGTGGCGAA
GGCGGCTCCCTGGTTCAACATTGACGCTGAGGTGCGAAAGNGTGGGTAGCAAACAG
GNTTAGATACCCTGGTAGTCCACGCNGTAAACGATGTCGACTAGCCGTTGGGTCCAT
TTAAGGGCTTAGTGGCGCATAAACGCGATAAGTCGACCGCCTGGGGAGTACGGCCG
CAAGGTTAAAACTCAAAGGAATTGACGGGGGCCCGCACAAGCGGTGGAGCATGTGGT
TTAATTCGATGCAACGCGAAAAACCTTACCAGCCCTTGACATCCTCGGAATCTTGCAG
AGATGTGAGAGTGCCTTCGGGAACCGAGAGACAGGTGCTGCATGGCTGTCGTCAGC
TCGTGTCGTGAGATGTTGGGTTAAGTCCCGTAACGAGCGCAACCCTTGTCCTTAGTT
GCCAGCGCGTCAAGGCGGGAACTCTAAGGAGACTGCCGGTGATAAACCGGAGGAAG
GTGGGGATGACGTCAAGTCATCATGGCCCTTATGGGCTGGGCTACACACGTGCTACA
ATGGCCGGTACAGAGCGTTGCGACCCCGCGAGGGTGAGCCAATCGCAGAAAACCGG
TCGTAGTCCGGATCGCAGTCTGCAACTCGACTGCGTGAAGTCGGAATCGCTAGTAAT
CGCGAATCAGCATGTCGSGGTGAATACGTTCCCGGGCCTTGTACACACCGCCSGTC
ACACCATGGGAGTTGGTTGCACCAGAAGTAGATCGCTTAACCGCAAGAMGGGCGTTT
ACCACGGTGTGTACACTGACTGGGGTGAAGTCGTACAAGG

2. **Something present in humans?:**

GCTAAACCTAGCCCCAAACCCACTCCACCTTACTACCAGACAACCTTAGCCAAACCA
TTTACCCAAATAAAGTATAGGCGATAGAAATTGAAACCTGGCGCAATAGATATAGTACC
GCAAGGGAAAGATGAAAAATTATAACCAAGCATAATATAGCAAGGACTAACCCCTATAC
CTTCTGCATAATGAATTAACTAGAAATAACTTTGCAAGGAGAGCCAAAGCTAAGACCC
CCGAAACCAGACGAGCTACCTAAGAACAGCTAAAAGAGCACACCCGTCTATGTAGCA
AAATAGTGGGAAGATTTATAGGTAGAGGCGACAAACCTACCGAGCCTGGTGATAGCTG
GTTGTCCAAGATAGAATCTTAGTTCAACTTTAAATTTGCCCACAGAACCCTCTAAATCC
CCTTGTAAATTTAACTGTTAGTCCAAAGAGGAACAGCTCTTTGGACACTAGGAAAAAAC
CTTGTAGAGAGAGTAAAAAATTTAACACCCATAGTAGGCCTAAAAGCAGCCACCAATTA
AGAAAGCGTTCAAGCTCAACACCCACTACCTAAAAAATCCCAAACATATAACTGAACT
CCTCACACCCAATTGGACCAATCTATCACCCTATAGAAGAACTAATGTTAGTATAAGTA
ACATGAAAACATTCTCCTCCGCATAAGCCTGCGTCAGATTAAAACACTGAACTGACAA
TTAACAGCCCAATATCTACAATCAACCAACAAGTCATTATTACCCTCACTGTCAACCC

	AACACAGGCATGCTCATAAGGAAAGGTTAAAAAAAGTAAAAGGAACTCGGCAAATCTTA CCCCGCCTGTTTACCAAAAACATCACCTCTAGCATCACCAGTATTAGAGGCACCGCC TGCCCAGTGACACATGTTTAACGGCCGCGGTACCCTAACCGTGCAAAGGTAGCATAA TCACTTGTTCCTTAAATAGGGACCTGTATGAATGGCTCCACGAGGGTTCAGCTGTCT CTTACTTTTAACCAGTGAAATTGACCTGCCCGTGAAGAGGCGGGCATAACACAGCAA GACGAGAAGACCCTATGGAGCTTTAATTTATTAATGCAAACAGTACCTAACAAACCCA CAGGTCCTAAACTACCAAACCTGCATTAAAAATTTCGGTTGGGGCGACCTCGGAGCA GAACCCAACCTCCGAGCAGTACATGCTAAGACTTCACCAGTCAAAGCGAACTACTATA CTCAATTGATCCAATAACTTGACCAACGGAACAAGTTACCCTAGGGATAACAGCGCAA TCCTATTCTAGAGTCCATATCAACAATAGGGTTTACGACCTCGATGTTGGATCAGGAC ATCCCGATGGTGCAGCCGCTATTAAAGGTTCGTTTGTTCAACGATTAAAGTCCTACGT GATCTGAGTTCAGACCGGAGTAATCCAGGTCGGTTTCTATCTACCTTCAAATTCCTCC CTGTACGAAAGGACAAGAGAAATAAGGCCTACTTCACAAAGCGCCTTCCCCCGTAAAT GATATCATCTCAACTTAGTATTATACCCACACCCACCCAAGAACAGGGTTT
Equipment	Internet work stations
Rules & Precautions	No microbial risks, only computer viruses ....
Experiences gained	• Familiarize yourself with a few theoretical and pratical aspects of computer-aided molecular sequence analysis techniques • Experience how the internet can be used to get phylogenetic information from a nucleotide sequence • Perform sequence database searches • Practice to download and understand sequence records from international data bases • Learn about the usefulness and limitations of some computer algorithms for sequence analysis
Timing	90 minutes
Reporting	Take notes on the exercise and present the results in your report and to the class.
Questions to be answered	1. What is the RDP? 2. What is an evolutionary distance (ED)? 3. How old is Earth? 4. How old are the oldest known microfossils? 5. Why are ribosomal RNA genes good evolutionary chronometers? 6. How does the universal tree support the idea that early Earth was very hot? 7. Describe why the results of phylogenetic community analyses of microbial habitats have been rather surprising. 8. Why are we using rRNA genes for studying phylogenetic relationships? 9. What effect do slightly different sequences have on the positioning in a phylogenetic group?