University of Zurich
Laboratory to Biology III - Diversity of Microorganisms
Assistant: Thomas Horath

Phylogeny of Prokaryotes based on 16S rRNA genes and an introduction to BLAST

Fabienne Häusler (fabienne@access.unizh.ch), Benjamin Homberger (s0373104@access.unizh.ch), Simone Stahel (s0370955@access.unizh.ch), Daniel Stocker (s0192499@access.unizh.ch), and Thomas Fahrni (thomi.f@access.unizh.ch)

Introduction

Without the potential of today's molecular biology and web-based sequence databases and tools, e.g. BLAST, early taxonomic classification of prokaryotes depended only on phenotypic characteristics such as shape, behaviour (chemo-, photo- or magnetotactic), membrane structure, metabolism, resistance to antibiotics, cell wall (Gram-positive or Gram-negative), pathogenicity/virulence and some other.

Some of these properties turned out to show the proximity of relationships between different microorganisms (e.g. Gram-staining), while others are taxonomy-independent, e.g. resistance to antibiotics, which can be transferred between different bacteria via plasmids.

The modern approach to phylogeny relies on molecular studies and sequence comparisons of genes and proteins. One of the most extensively used method to develop the phylogeny of both prokaryotes and eukaryotes was pioneered in the early 1970s by Carl Woese. This so-called "SSU sequencing" or "16S/18S sequencing" is based on the 16S (prokaryotes) and 18S (eukaryotes) ribosomal RNA (rRNA) genes.

16S/18S rRNA

All living organisms contain the small (16S or 18S) and the large (23S or 28S) subunit rRNA. Since these subunits are essential for protein synthesis, they all have the same function and must have been developed in the early stages of life. Mutations in these genes can affect directly the ordinary functioning of the ribosome and thus, only minor changes in these genes are allowed. Otherwise, the ribosome can lose its function resulting in the elimination of mutated organisms. Since 16S rRNA is rather sensitive to mutations, the corresponding gene seems to contain a large number of highly conserved regions. Some of them don't affect the ribosome's function and mutations can accumulate over evolutionary times.

Given the aspects above, 16S rRNA, as well as 18S rRNA, is a useful evolutionary chronometer to estimate the relationship of organisms.

BLAST

Molecular sequence data gathered worldwide is collected and made available by public databases such as NCBI (National Center for Biotechnology Information). Bioinformatics provides software tools to deal with the vast amount of data (over 36 billions base pairs and 30 millions sequences alone at GenBank). One of this tool is BLAST (Basic Local Alignment Search Tool), which searches for patches of similarity (not identity!) using similarity scoring matrices.

blastn is used to compare a DNA sequence with DNA sequences in the database. It returns a list of best matches with their bit score and E-value.

The bit score S' is a normalized (scoring matrix independent) measure for similarity and is calculated from the raw score S with

$$S' = \frac{\lambda \cdot S - \ln(K)}{\ln(2)}$$

where λ is a scoring matrix specific constant and K is a minor correction constant.

The E-value indicates the significance of a match (the probability that the matching sequence matches not only by chance) and is calculated with

$$E = m\,n\,2^{-S'}$$

where m is the length of database entries and n is the length of the search sequence.

The tree of life

To render the tree of life based on sequence data, the computer calculates the evolutionary distance (E_D) between two sequences:

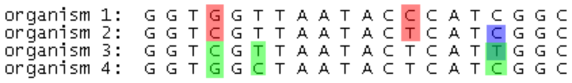
$$E_D = \frac{non-identical/non-similar\ nucleotides}{sequence\ length}$$

Note that "non-similar nucleotides" means that the mutated nucleotide has an effect on the amino acid sequence.

Given four organisms with their respective sequence (20 nucleotides) as shown in figure 1, we get the following E_D:

Organisms	Mutations	Seq. length	E_D[1]
1 > 2	2	20	0.1
1 > 3	3	20	0.15
1 > 4	2	20	0.1
2 > 3	1	20	0.05
2 > 4	2	20	0.1
3 > 4	3	20	0.15

Table 1: Calculated E_D for given sequences



```
organism 1:  G G T G G T T A A T A C C C A T C G G C
organism 2:  G G T C G T T A A T A C T C A T C G G C
organism 3:  G G T C G T T A A T A C T C A T T G G C
organism 4:  G G T G G C T A A T A C T C A T C G G C
```

Fig. 1: Four sample sequences from which E_D is calculated.

Assuming that the given organisms emerged in the order 1 > 2 > 3 > 4, the calculated E_D*'s* in Table 1 suggest that organism 4 (E_D = 0.1) is nearer related to organism 1 than organism 3 (E_D = 0.15) to organism 1, even though organism 4 is a descendant of organism 3. This discrepancy is related to the fact, that we didn't consider that mutated nucleotides can mutate back (A>C>A) or mutate again (A>C>T). To eliminate this difference, the E_D is corrected with a statistical parameter which includes the mutation-rate and the probability that a back- or forward-mutation occurs.

With the corrected E_D it's possible to draw the most probable phylogenetic tree, where the length of the branches is proportional to the number of mutations (Fig. 2).

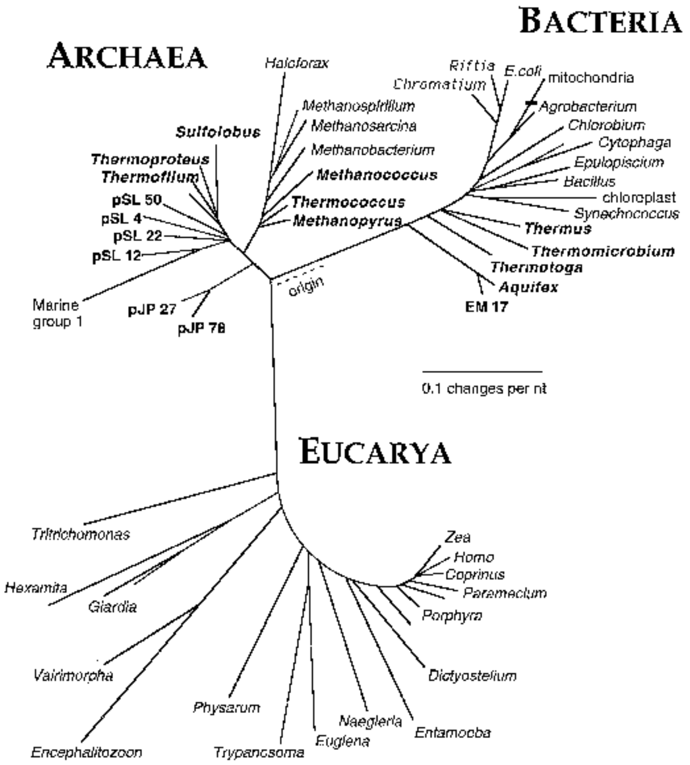


Fig. 2: Phylogenetic tree of life

[1] In this example, E_D is based on identity, not similarity.

Exercices

What are the sequences available in BLAST that share the highest relative level of similarity with the sequences given (see Appendix)? To which organisms do sequences 1 and 2 belong?

The most similar sequence to sequence 1 is gi|3881927|emb|Y12376.1|COY12376, the *Chromatium okenii* 16S rRNA gene. Several matches for sequence 2 have equal bit scores and E-values, e.g. gi|1944628|gb|J01415.1|HUMMTCG, which is the complete genome of the human mitochondrion.

What is the minimal length of the nucleotide sequence to be typed into the BLAST window to still get the correct nearest organisms? Is this length always the same or does it depend on the fragment location which you choose from the entire sequence?

It's impossible to specify a general minimal length for the search sequence. In general, the longer the fragment, the better the match: Short sequences don't produce significant matches (see Table 2). The results depend on the fragment location as well: In highly conserved regions (e.g. location 1020), the fragment has to be much longer (>170) than in more variable regions (e.g. location 1) where only <58 nucleotides are needed.

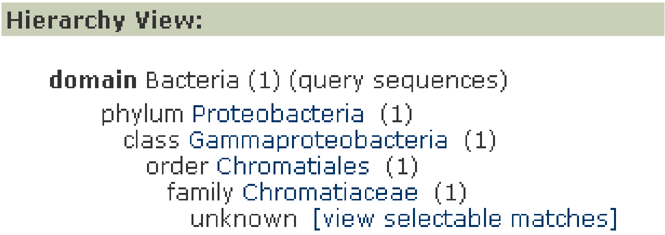
Location of first nucleotide in seq.	**Fragment length**	**Nearest organism (blastn)**
1	8	(sequence too short, no significant matches)
1	16	*Chromatium okenii* / *Anabaena circinalis*
1	33	*Chromatium okenii* / *Anabaena circinalis*
1	57	*Chromatium okenii*
1020	57	*Pseudomonas* sp.
1020	62	*Alcanivorax* sp.
1020	70	*Chromatium okenii* / Phototrophic bacteria
1020	113	*Chromatium okenii* / Phototrophic bacteria
1020	170	*Chromatium okenii* / Phototrophic bacteria

Table 2: Nearest organism depending on fragment location and length
(Used sequence: Chromatium okenii 16S rRNA gene)

Take a small fragment of your favourite sequence and exchange one or more nucleotides in it. Does this new sequence still have the same nearest relatives?

The sequence used is the "1-57" fragment from sequence 1 (cf. Table 2), because it clearly belongs to *Chromatium okenii*. A single random mutation has no effect on the nearest organism, but 10 random mutations lead to very different nearest organisms, e.g. *G. wittrockiana*, *Bathymodiolus* sp., *Nostoc* sp. and others.

On the Ribosomal Database Project it is possible to create sequence based trees. Can we fit the inhabitant of lake Cadagno (sequence 1) into its adequate group?



Hierarchy View:

domain Bacteria (1) (query sequences)
phylum Proteobacteria (1)
class Gammaproteobacteria (1)
order Chromatiales (1)
family Chromatiaceae (1)
unknown [view selectable matches]

Fig. 3: Result of RDP sequence match. Note that the link [view selectable matches] leads to Chromatium okenii 16S rRNA gene.

Literature/Links

Madigan, M. T., J. M. Martinko and J. Parker: "Brock - Biology of Microorganisms", 10th edition, 2003. Prentice Hall

GenBank Statistics ("GenBank Growth"):
http://www.ncbi.nlm.nih.gov/Genbank/genbankstats.html

NCBI BLAST:
http://www.ncbi.nlm.nih.gov/BLAST/

NCBI Handbook:
http://http://www.ncbi.nlm.nih.gov/books/bv.fcgi?rid=handbook

Ribosomal Database Project (RDP):
http://rdp.cme.msu.edu/

Appendix

Sequence 1: A famous inhabitant of lake Cadagno

```
CGTGGCGGTATGCTTAACACATGCAAGTCGAACGTCAAAGGTCTTCGGATTGAGTAGCGTGGCGGACGGGTGAGTAAAGCGTGGGAATCTGCCTTGCAGT
GGGGGATAACCCGGGGAAACTCGGGCTAATACCGCATACGCCCTACGGGGGAAAGGGGGCTTTGGCTCTCGTTGCAAGATGAGCCCACGTCCGATTAGCT
AGTTGGTAGGGTAAAGGCCTACCAAGGCGACGATSGGTCGCTGGTCTGAGAGGATGACCAGCCACACTGGGACTGAGACACGGCCCAGACTCCTACGGGA
GGCAGCAGTGGGAATATTGGACAATGGGGGAAACCCTGATCCAGCAATACCGCGTGTGTGAAGAAGGCCTGCGGGTTGTAAAGCACTTTCAGTGGAAAG
AAAACCTGGTGGTTAATACCCATCGGCTTTGACGTTACTCACAAAAGAAGCACCGGCTAACTCCGTGCCAGCAGCCGCGGTAATACGGAGGGTGCAAGCG
TTAATCGGAATTACTGGGCGTAAAGCGCACGTAGGCGGCGCCGTCAGTCCGATGTGAAAGCCCTGGGCTTAACCTGGGAACTGCATTGGATACTGCGGCG
CTAGAATGTGAAAGAGGGGAGTGGAATTCCAGGTGTAGCGGTGAAATGCGTAGAGATCTGGAGGAACACCAGTGGCGAAGGCGGCTCCCTGGTTCAACAT
TGACGCTGAGGTGCGAAAGNGTGGGTAGCAAACAGGNTTAGATACCCTGGTAGTCCACGCNGTAAACGATGTCGACTAGCCGTTGGGTCCATTTAAGGGC
TTAGTGGCGCATAAACGCGATAAGTCGACCGCCTGGGGAGTACGGCCGCAAGGTTAAAACTCAAAGGAATTGACGGGGGCCCGCACAAGCGGTGGAGCAT
GTGGTTTAATTCGATGCAACGCGAAAAACCTTACCAGCCCTTGACATCCTCGGAATCTTGCAGAGATGTGAGAGTGCCTTCGGGAACCGAGAGACAGGTG
CTGCATGGCTGTCGTCAGCTCGTGTCGTGAGATGTTGGGTTAAGTCCCGTAACGAGCGCAACCCTTGTCCTTAGTTGCCAGCGCGTCAAGGCGGAACTC
TAAGGAGACTGCCGGTGATAAACCGGAGGAAGGTGGGGATGACGTCAAGTCATCATGGCCCTTATGGGCTGGGCTACACACGTGCTACAATGGCCGGTAC
AGAGCGTTGCGACCCCGCGAGGGTGAGCCAATCGCAGAAAACCGGTCGTAGTCCGGATCGCAGTCTGCAACTCGACTGCGTGAAGTCGGAATCGCTAGTA
ATCGCGAATCAGCATGTCGSGGTGAATACGTTCCCGGGCCTTGTACACACCGCCSGTCACACCATGGGAGTTGGTTGCACCAGAAGTAGATCGCTTAACC
GCAAGAMGGGCGTTTACCACGGTGTGTACACTGACTGGGGTGAAGTCGTACAAGG
```

Sequence 2: Something present in humans?

```
GCTAAACCTAGCCCCAAACCCACTCCACCTTACTACCAGACAACCTTAGCCAAACCATTTACCCAAATAAAGTATAGGCGATAGAAATTGAAACCTGGCG
CAATAGATATAGTACCGCAAGGGAAAGATGAAAAATTATAACCAAGCATAATATAGCAAGGACTAACCCCTATACCTTCTGCATAATGAATTAACTAGAA
ATAACTTTGCAAGGAGAGCCAAAGCTAAGACCCCCGAAACCAGACGAGCTACCTAAGAACAGCTAAAAGAGCACACCCGTCTATGTAGCAAAATAGTGGG
AAGATTTATAGGTAGAGGCGACAAACCTACCGAGCCTGGTGATAGCTGGTTGTCCAAGATAGAATCTTAGTTCAACTTTAAATTTGCCCACAGAACCCTC
TAAATCCCCTTGTAAATTTAACTGTTAGTCCAAAGAGGAACAGCTCTTTGGACACTAGGAAAAAACCTTGTAGAGAGAGTAAAAAATTTAACACCCATAG
TAGGCCTAAAAGCAGCCACCAATTAAGAAAGCGTTCAAGCTCAACACCCACTACCTAAAAAATCCCAAACATATAACTGAACTCCTCACACCCAATTGGA
CCAATCTATCACCCTATAGAAGAACTAATGTTAGTATAAGTAACATGAAAACATTCTCCTCCGCATAAGCCTGCGTCAGATTAAAACACTGAACTGACAA
TTAACAGCCCAATATCTACAATCAACCAACAAGTCATTATTACCCTCACTGTCAACCCAACACAGGCATGCTCATAAGGAAAGGTTAAAAAAAGTAAAAG
GAACTCGGCAAATCTTACCCCGCCTGTTTACCAAAAACATCACCTCTAGCATCACCAGTATTAGAGGCACCGCCTGCCCAGTGACACATGTTTAACGGCC
GCGGTACCCTAACCGTGCAAAGGTAGCATAATCACTTGTTCCTTAAATAGGGACCTGTATGAATGGCTCCACGAGGGTTCAGCTGTCTCTTACTTTTAAC
CAGTGAAATTGACCTGCCCGTGAAGAGGCGGGCATAACACAGCAAGACGAGAAGACCCTATGGAGCTTTAATTTATTAATGCAAACAGTACCTAACAAAC
CCACAGGTCCTAAACTACCAAACCTGCATTAAAAATTTCGGTTGGGGCGACCTCGGAGCAGAACCCAACCTCCGAGCAGTACATGCTAAGACTTCACCAG
TCAAAGCGAACTACTATACTCAATTGATCCAATAACTTGACCAACGGAACAAGTTACCCTAGGGATAACAGCGCAATCCTATTCTAGAGTCCATATCAAC
AATAGGGTTTACGACCTCGATGTTGGATCAGGACATCCCGATGGTGCAGCCGCTATTAAAGGTTCGTTTGTTCAACGATTAAAGTCCTACGTGATCTGAG
TTCAGACCGGAGTAATCCAGGTCGGTTTCTATCTACCTTCAAATTCCTCCCTGTACGAAAGGACAAGAGAAATAAGGCCTACTTCACAAAGCGCCTTCCC
CCGTAAATGATATCATCTCAACTTAGTATTATACCCACACCCACCCAAGAACAGGGTTT
```