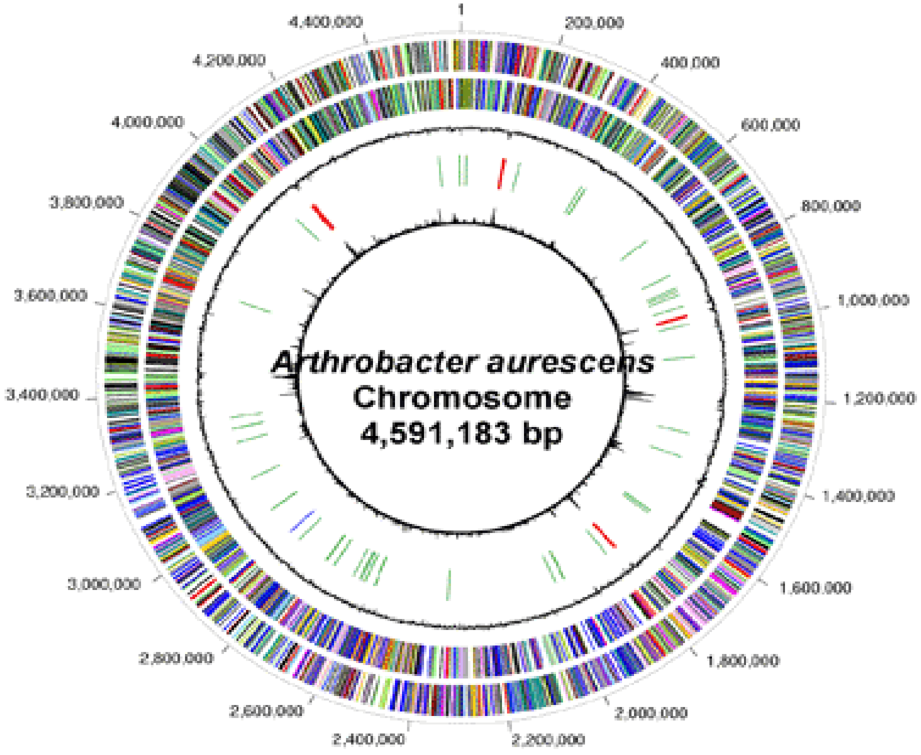
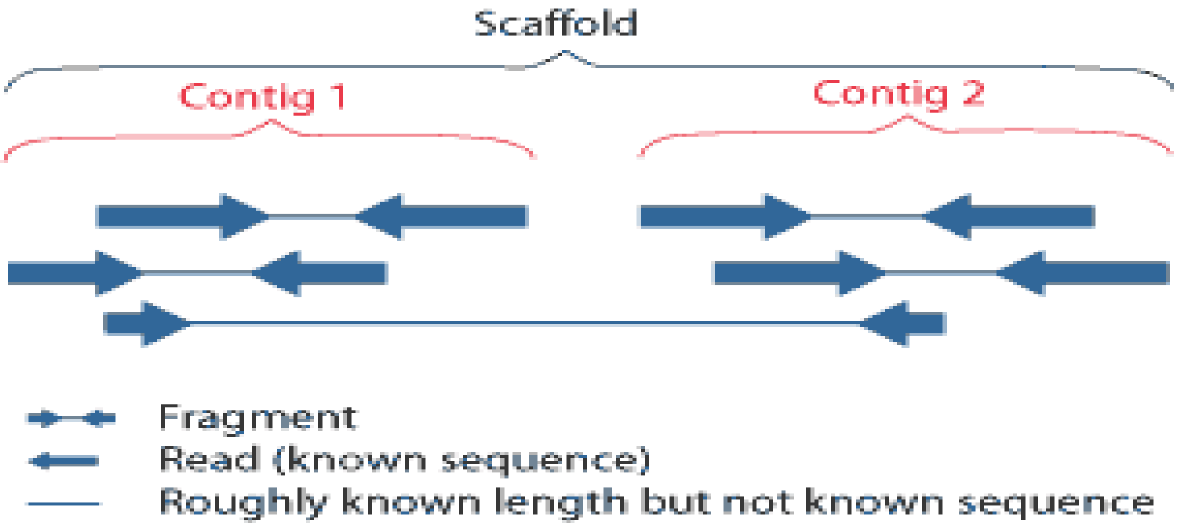
Genome Practical
Microbial Genomics



What Time Is It? ....Its Genome Time!

Sequence data

- Contigs: overlapping DNA fragments (sequence reads) that form a consensus region of DNA
- Paired-end sequencing

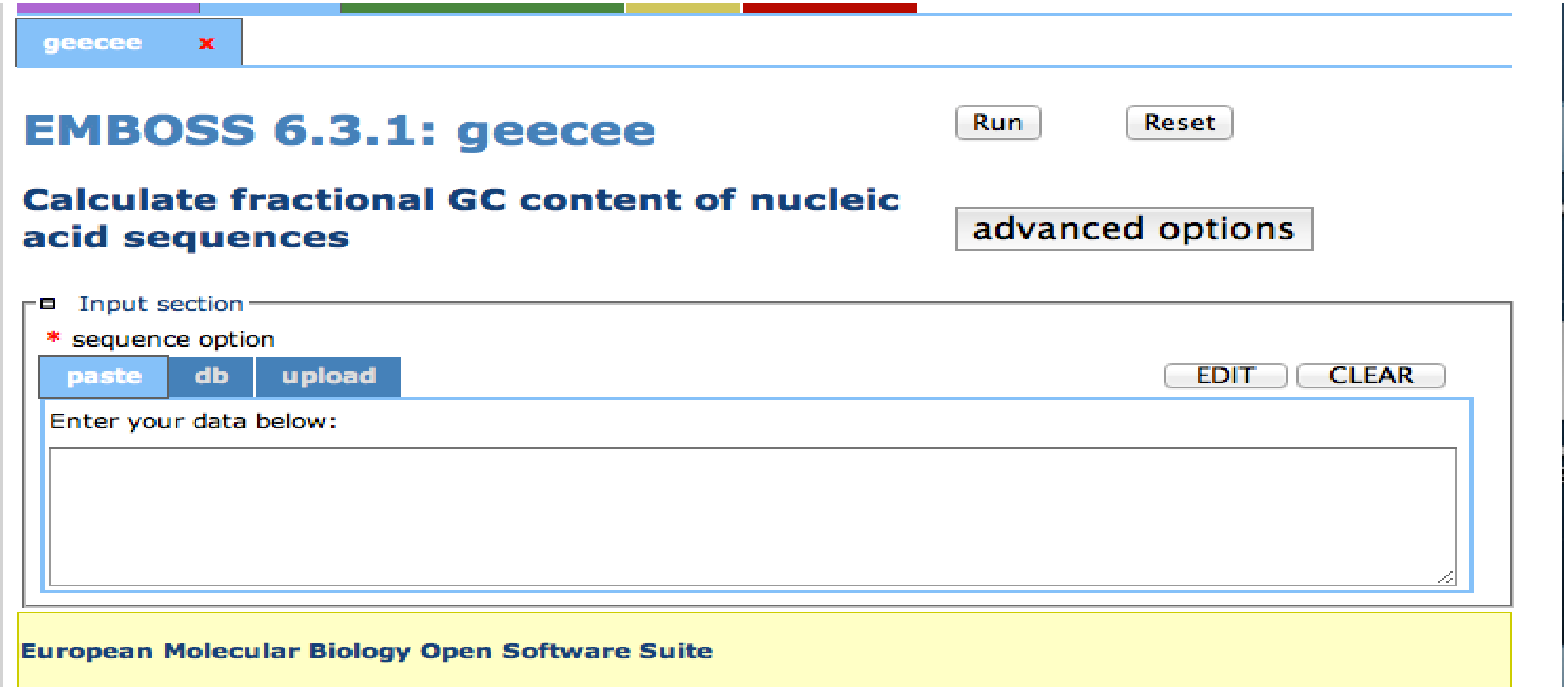


Analyzing GC Content

- Whole genomes can be distinguished by GC content in some cases
- These differences can arise via lateral gene transfer

Exercise 1

- Paste the sequence data for each genome respectively
- Save the output



EMBOSS 6.3.1: geecee

Calculate fractional GC content of nucleic acid sequences

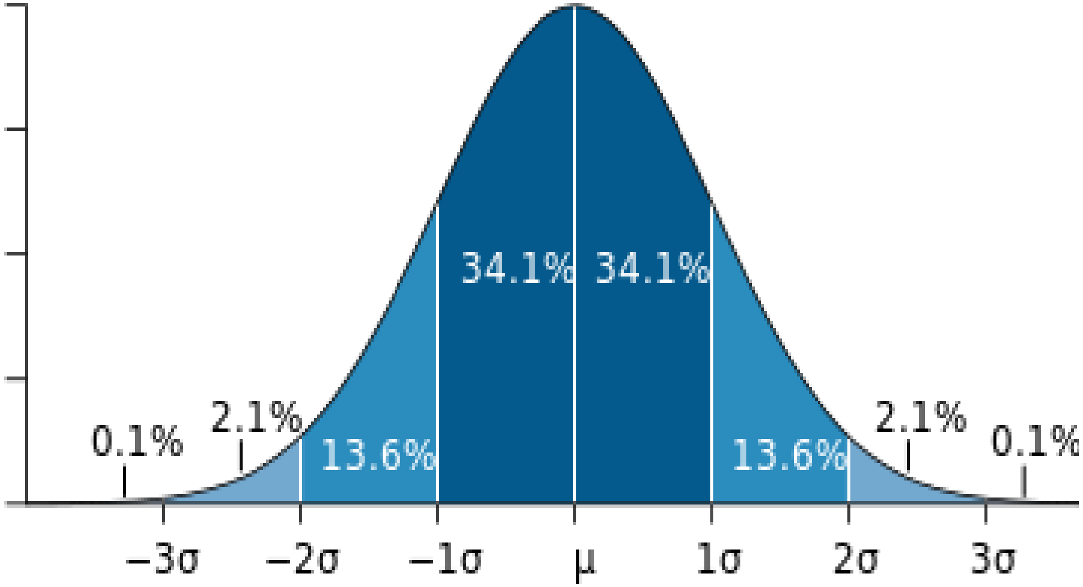
* sequence option

Enter your data below:

European Molecular Biology Open Software Suite

Ex1 cont’d

- Normally distributed data
- 95% - 2 stdev
- Determine the mean(μ)
- Determine standard deviation (σ)



$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}, \quad \text{where} \quad \mu = \frac{1}{N}\sum_{i=1}^{N} x_i.$$

```
from sequence import readFastaFile
def countGC(sequence):
    Gs=sequence.count('G')
    Cs=sequence.count('C')
    per=float(Gs+Cs)/len(sequence)
    return round(per,2)

def meanGC(GCcounts):

mean=float(sum(GCcounts))/len(GCcounts)
    return round(mean,2)

def stdev(values,mean):
    """This is NOT the sample standard
deviation """
    vals=[]
    for i in range(len(values)):
        vals.append((values[i]-mean)**2)

stdev=math.sqrt((1/float(len(values)))*sum(
vals))
```

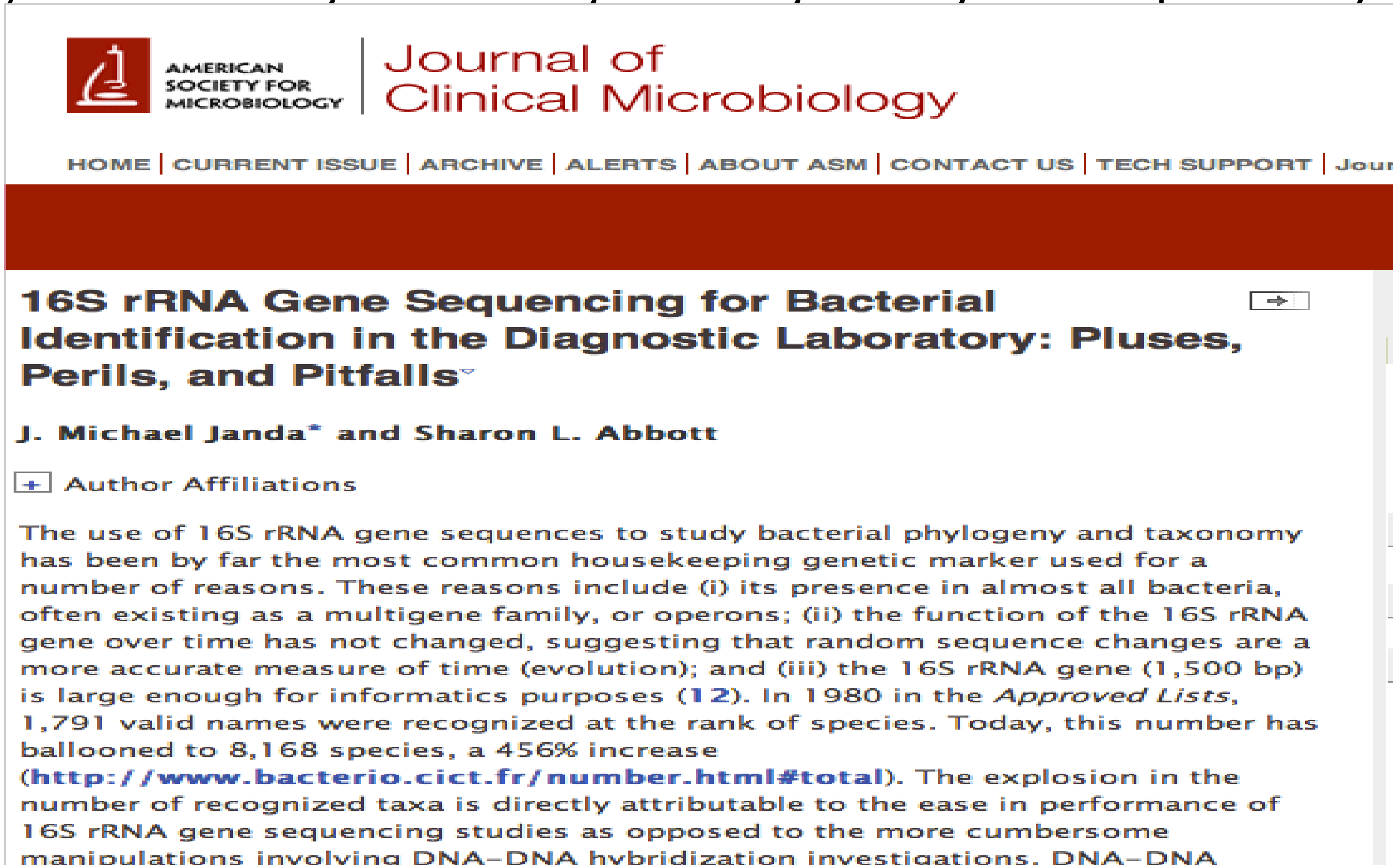
```
#read in fasta files
seqs = readFastaFile('Genome1.fasta'
seqs2 = readFastaFile('Genome2.fasta
seqs3 = readFastaFile('Genome3.fasta

#Count GC in each contig
counts=[]
for seq in seqs: #change seqs
    count=countGC(seq)
    counts.append(count)

#(Repeat for other genomes)
print "GENOME1 Processing"
G1mean= meanGC(counts)
print "mean",G1mean
std= stdev(counts,G1mean)
print "stdev",std
upper=G1mean+2*std
print "upper",round(upper,2)
lower=G1mean-2*std
print "lower",round(lower,2)
```

Table Exercise 1

Genomes	Mean	Standard Dev.	Upper limits	Lower limits
Genome1	Mean1	Stdev1	Upper1	Lower1
Genome2	Mean2	Stdev2	Upper2	Lower2
Genome3	Mean3	Stdev3	Upper3	Lower3

Taxonomic Identification (16S rRNA)

- 1) Present in all bacteria as a multigene family
- 2) The function of 16S rRNA genes have not changed over time (changes in sequence) can accurately determine evolution
- 3) 16S rRNA gene is large enough for gene sequencing



AMERICAN SOCIETY FOR MICROBIOLOGY | Journal of Clinical Microbiology

HOME | CURRENT ISSUE | ARCHIVE | ALERTS | ABOUT ASM | CONTACT US | TECH SUPPORT | Jour

16S rRNA Gene Sequencing for Bacterial Identification in the Diagnostic Laboratory: Pluses, Perils, and Pitfalls

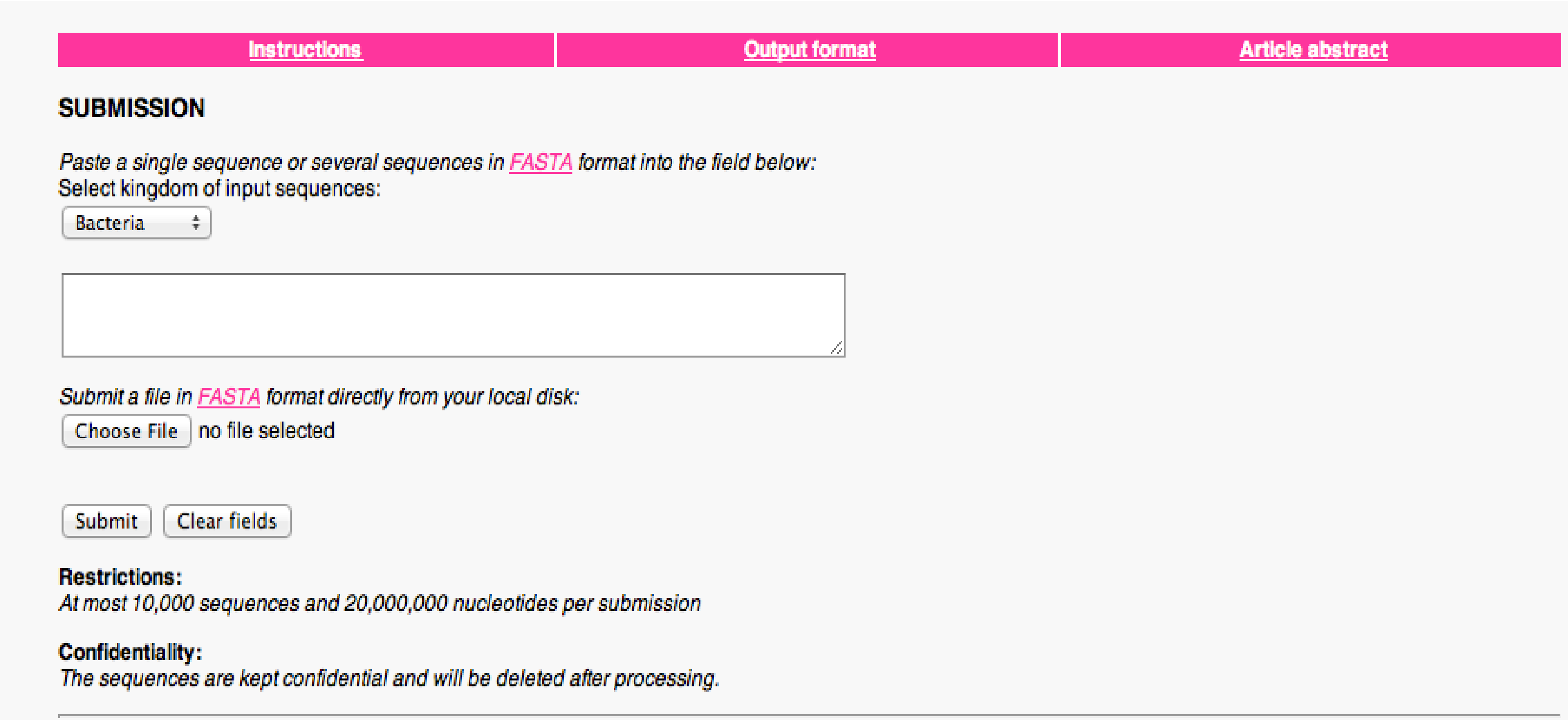
J. Michael Janda* and Sharon L. Abbott

⊞ Author Affiliations

The use of 16S rRNA gene sequences to study bacterial phylogeny and taxonomy has been by far the most common housekeeping genetic marker used for a number of reasons. These reasons include (i) its presence in almost all bacteria, often existing as a multigene family, or operons; (ii) the function of the 16S rRNA gene over time has not changed, suggesting that random sequence changes are a more accurate measure of time (evolution); and (iii) the 16S rRNA gene (1,500 bp) is large enough for informatics purposes (12). In 1980 in the *Approved Lists*, 1,791 valid names were recognized at the rank of species. Today, this number has ballooned to 8,168 species, a 456% increase (http://www.bacterio.cict.fr/number.html#total). The explosion in the number of recognized taxa is directly attributable to the ease in performance of 16S rRNA gene sequencing studies as opposed to the more cumbersome manipulations involving DNA-DNA hybridization investigations. DNA-DNA

Exercise 2

- RNAmmer Creates a HMM from the structural alignments to predict rRNA genes



SUBMISSION

Paste a single sequence or several sequences in FASTA format into the field below:

Select kingdom of input sequences: Bacteria

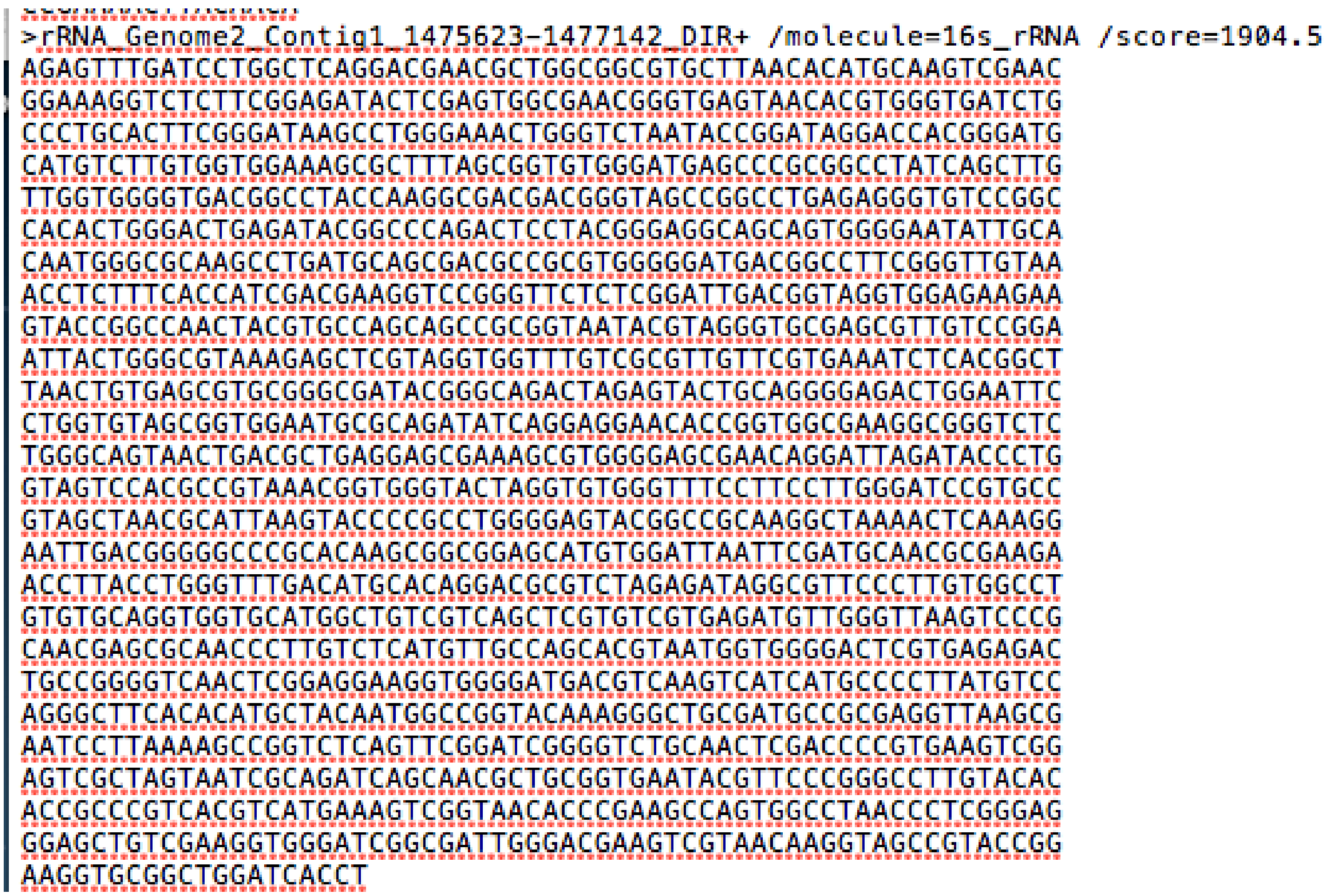
Submit a file in FASTA format directly from your local disk:

Restrictions:
At most 10,000 sequences and 20,000,000 nucleotides per submission

Confidentiality:
The sequences are kept confidential and will be deleted after processing.

Exercise 2 cont'd

- Green genes → Compare → BLAST



```
>rRNA_Genome2_Contig1_1475623-1477142_DIR+ /molecule=16s_rRNA /score=1904.5
AGAGTTTGATCCTGGCTCAGGACGAACGCTGGCGGCGTGCTTAACACATGCAAGTCGAAC
GGAAAGGTCTCTTCGGAGATACTCGAGTGGCGAACGGGTGAGTAACACGTGGGTGATCTG
CCCTGCACTTCGGGATAAGCCTGGGAAACTGGGTCTAATACCGGATAGGACCACGGGATG
CATGTCTTGTGGTGGAAAGCGCTTTAGCGGTGTGGGATGAGCCCGCGGCCTATCAGCTTG
TTGGTGGGGTGACGGCCTACCAAGGCGACGACGGGTAGCCGGCCTGAGAGGGTGTCCGGC
CACACTGGGACTGAGATACGGCCCAGACTCCTACGGGAGGCAGCAGTGGGGAATATTGCA
CAATGGGCGCAAGCCTGATGCAGCGACGCCGCGTGGGGGATGACGGCCTTCGGGTTGTAA
ACCTCTTTCACCATCGACGAAGGTCCGGGTTCTCTCGGATTGACGGTAGGTGGAGAAGAA
GTACCGGCCAACTACGTGCCAGCAGCCGCGGTAATACGTAGGGTGCGAGCGTTGTCCGGA
ATTACTGGGCGTAAAGAGCTCGTAGGTGGTTTGTCGCGTTGTTCGTGAAATCTCACGGCT
TAACTGTGAGCGTGCGGGCGATACGGGCAGACTAGAGTACTGCAGGGGAGACTGGAATTC
CTGGTGTAGCGGTGGAATGCGCAGATATCAGGAGGAACACCGGTGGCGAAGGCGGGTCTC
TGGGCAGTAACTGACGCTGAGGAGCGAAAGCGTGGGGAGCGAACAGGATTAGATACCCTG
GTAGTCCACGCCGTAAACGGTGGGTACTAGGTGTGGGTTTCCTTCCTTGGGATCCGTGCC
GTAGCTAACGCATTAAGTACCCCGCCTGGGGAGTACGGCCGCAAGGCTAAAACTCAAAGG
AATTGACGGGGGCCCGCACAAGCGGCGGAGCATGTGGATTAATTCGATGCAACGCGAAGA
ACCTTACCTGGGTTTGACATGCACAGGACGCGTCTAGAGATAGGCGTTCCCTTGTGGCCT
GTGTGCAGGTGGTGCATGGCTGTCGTCAGCTCGTGTCGTGAGATGTTGGGTTAAGTCCCG
CAACGAGCGCAACCCTTGTCTCATGTTGCCAGCACGTAATGGTGGGGACTCGTGAGAGAC
TGCCGGGGTCAACTCGGAGGAAGGTGGGGATGACGTCAAGTCATCATGCCCCTTATGTCC
AGGGCTTCACACATGCTACAATGGCCGGTACAAAGGGCTGCGATGCCGCGAGGTTAAGCG
AATCCTTAAAAGCCGGTCTCAGTTCGGATCGGGGTCTGCAACTCGACCCCGTGAAGTCGG
AGTCGCTAGTAATCGCAGATCAGCAACGCTGCGGTGAATACGTTCCCGGGCCTTGTACAC
ACCGCCCGTCACGTCATGAAAGTCGGTAACACCCGAAGCCAGTGGCCTAACCCTCGGGAG
GGAGCTGTCGAAGGTGGGATCGGCGATTGGGACGAAGTCGTAACAAGGTAGCCGTACCGG
AAGGTGCGGCTGGATCACCT
```

Exercise 3

- The reading frame
- Why? (think about how we select the best reading frame)

Exercise 4

- What is the protein (hypothetical proteins are acceptable)
- What organism is it from

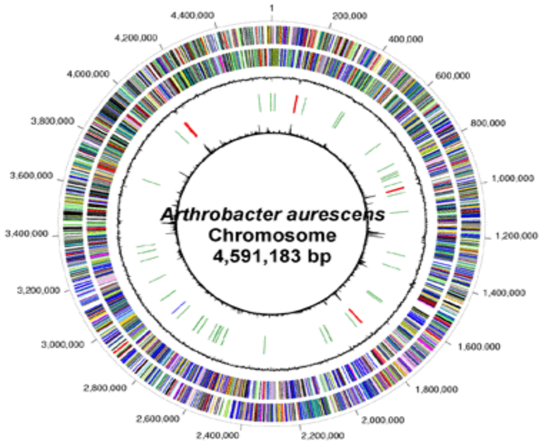
Exercise 5

- A) common pathways
- B)
- Think about the organism (day to day activities, energy requirements)
- Link pathways to key factors that allow the organism to thrive

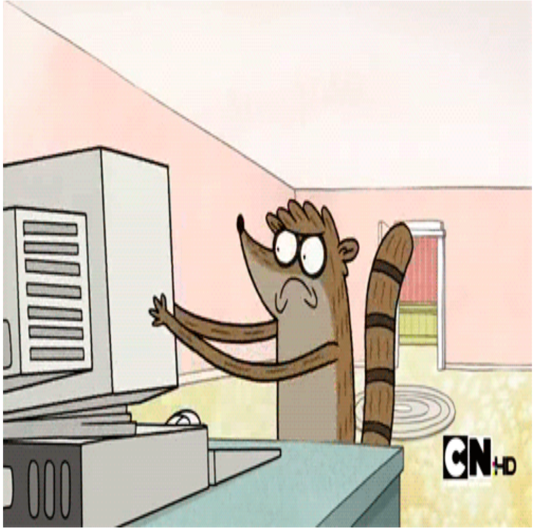
Exercise 6

- Look up each organism
- Noting differences
- Consider the pathways that are *different* between the organisms;
- Link the pathways to the differences
- Use **references** to scientific journals

Genome Practical
Microbial Genomics

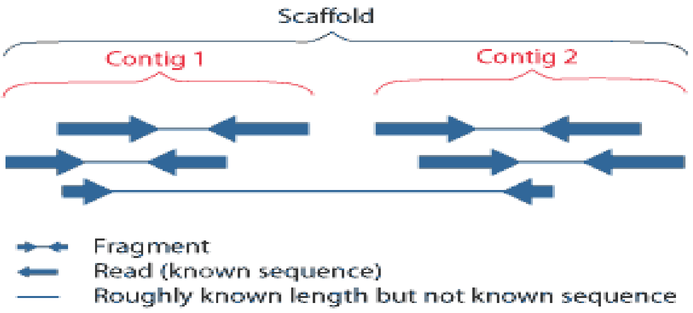


What Time Is It? ....Its Genome Time!



Sequence data

- Contigs: overlapping DNA fragments (sequence reads) that form a consensus region of DNA
- Paired-end sequencing

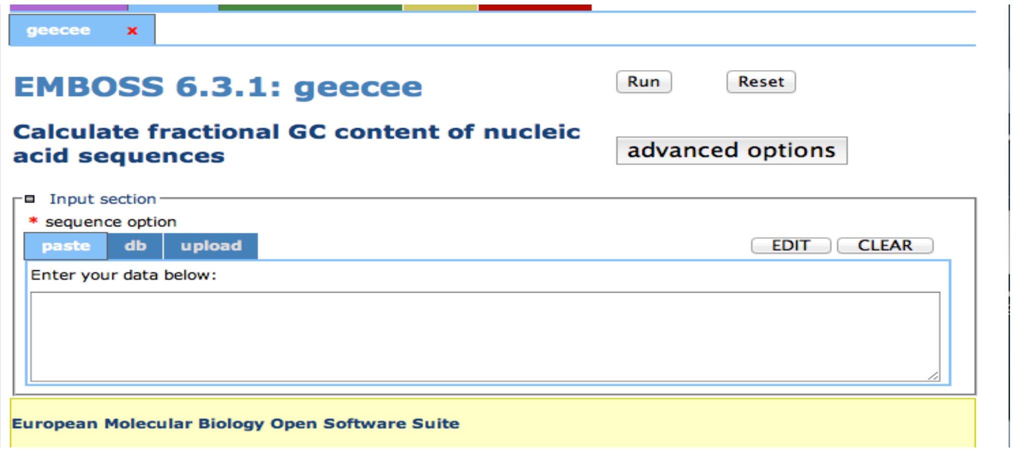


Analyzing GC Content

- Whole genomes can be distinguished by GC content in some cases
- These differences can arise via lateral gene transfer

Exercise 1

- Paste the sequence data for each genome respectively
- Save the output



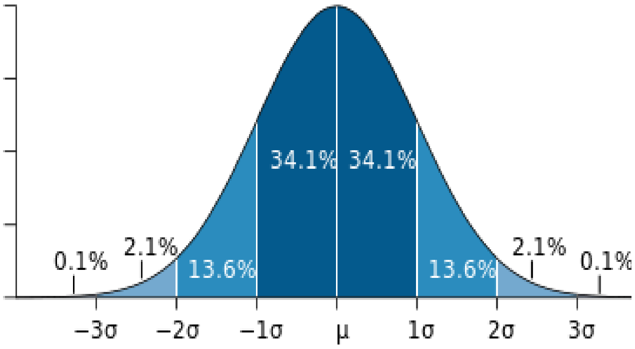
EMBOSS 6.3.1: geecee

Calculate fractional GC content of nucleic acid sequences

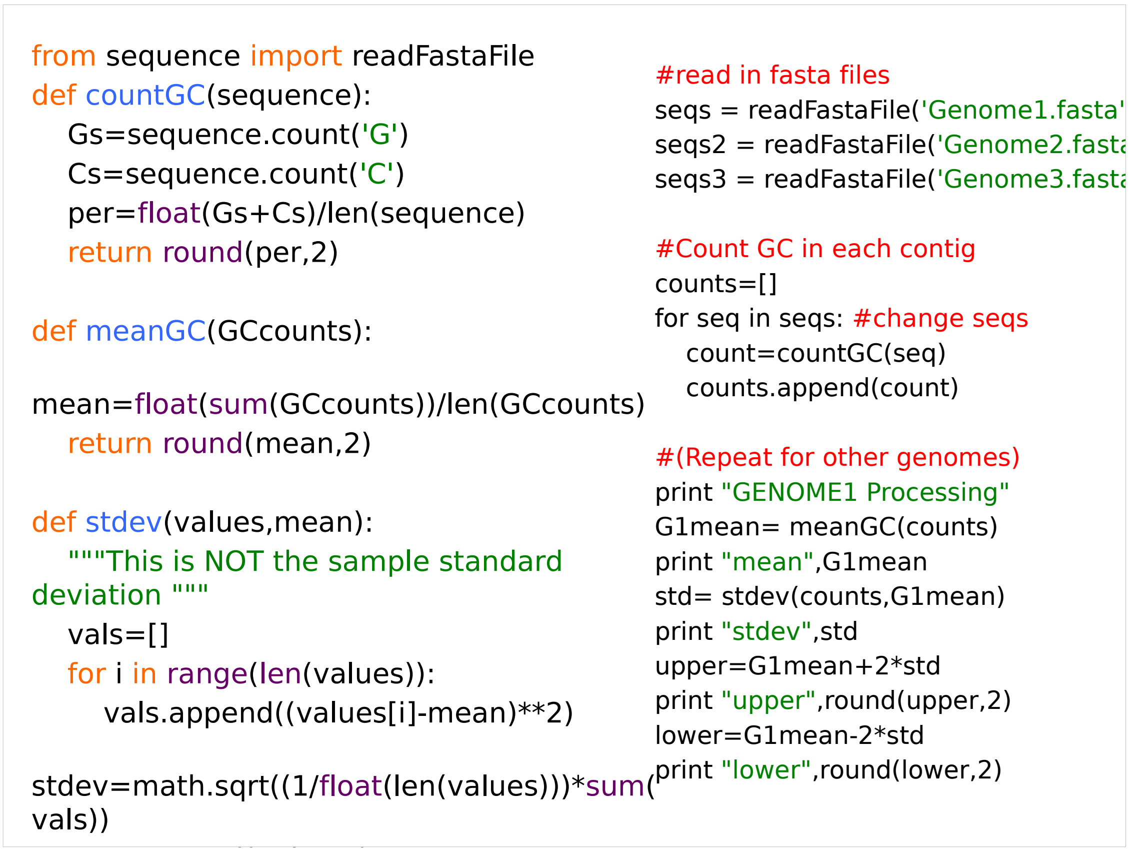
European Molecular Biology Open Software Suite

Ex1 cont'd

- Normally distributed data
- 95% - 2 stdev
- Determine the mean(μ)
- Determine standard deviation (σ)



$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}, \quad \text{where} \quad \mu = \frac{1}{N}\sum_{i=1}^{N} x_i.$$



```
from sequence import readFastaFile
def countGC(sequence):
    Gs=sequence.count('G')
    Cs=sequence.count('C')
    per=float(Gs+Cs)/len(sequence)
    return round(per,2)

def meanGC(GCcounts):

mean=float(sum(GCcounts))/len(GCcounts)
    return round(mean,2)

def stdev(values,mean):
    """This is NOT the sample standard
deviation """
    vals=[]
    for i in range(len(values)):
        vals.append((values[i]-mean)**2)

stdev=math.sqrt((1/float(len(values)))*sum(
vals))
```

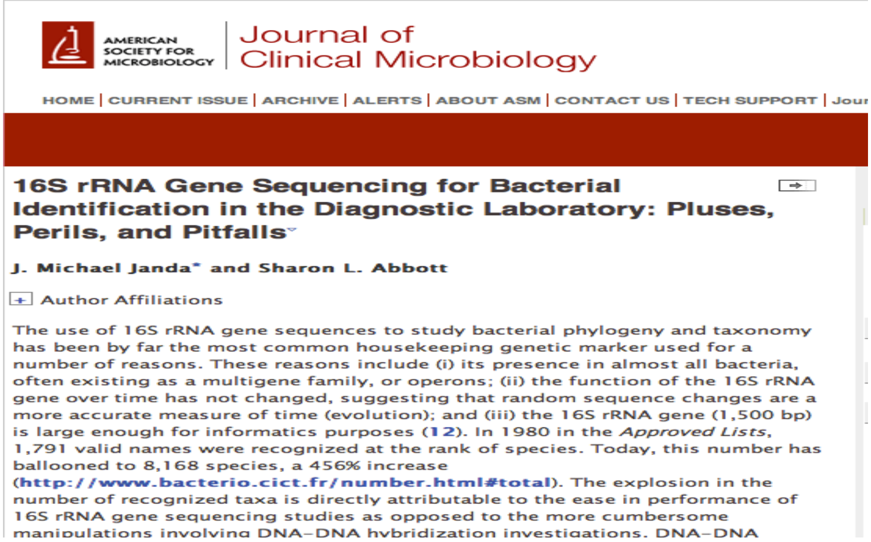
```
#read in fasta files
seqs = readFastaFile('Genome1.fasta'
seqs2 = readFastaFile('Genome2.fasta
seqs3 = readFastaFile('Genome3.fasta

#Count GC in each contig
counts=[]
for seq in seqs: #change seqs
    count=countGC(seq)
    counts.append(count)

#(Repeat for other genomes)
print "GENOME1 Processing"
G1mean= meanGC(counts)
print "mean",G1mean
std= stdev(counts,G1mean)
print "stdev",std
upper=G1mean+2*std
print "upper",round(upper,2)
lower=G1mean-2*std
print "lower",round(lower,2)
```

Table Exercise 1

Genomes	Mean	Standard Dev.	Upper limits	Lower limits
Genome1	Mean1	Stdev1	Upper1	Lower1
Genome2	Mean2	Stdev2	Upper2	Lower2
Genome3	Mean3	Stdev3	Upper3	Lower3

Taxonomic identification (16S rRNA)

- 1) Present in all bacteria as a multigene family
- 2) The function of 16S rRNA genes have not changed over time (changes in sequence) can accurately determine evolution
- 3) 16S rRNA gene is large enough for gene sequencing



AMERICAN SOCIETY FOR MICROBIOLOGY | Journal of Clinical Microbiology

HOME | CURRENT ISSUE | ARCHIVE | ALERTS | ABOUT ASM | CONTACT US | TECH SUPPORT | Jour

16S rRNA Gene Sequencing for Bacterial Identification in the Diagnostic Laboratory: Pluses, Perils, and Pitfalls

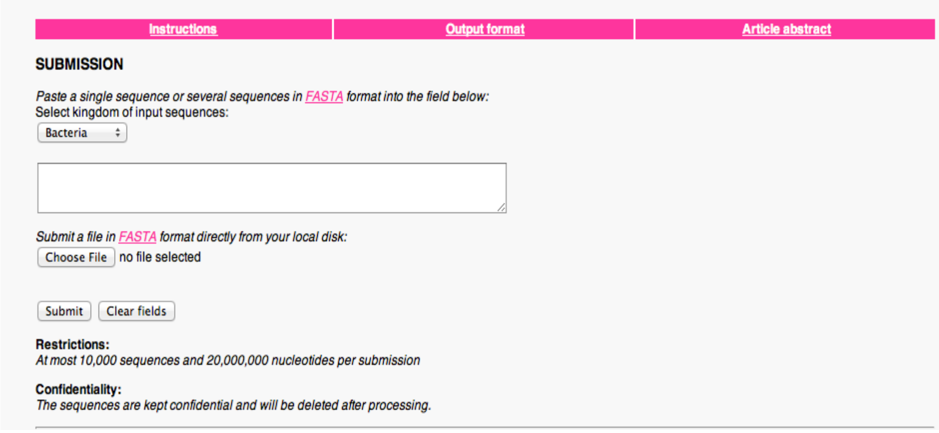
J. Michael Janda* and Sharon L. Abbott

Author Affiliations

The use of 16S rRNA gene sequences to study bacterial phylogeny and taxonomy has been by far the most common housekeeping genetic marker used for a number of reasons. These reasons include (i) its presence in almost all bacteria, often existing as a multigene family, or operons; (ii) the function of the 16S rRNA gene over time has not changed, suggesting that random sequence changes are a more accurate measure of time (evolution); and (iii) the 16S rRNA gene (1,500 bp) is large enough for informatics purposes (12). In 1980 in the *Approved Lists*, 1,791 valid names were recognized at the rank of species. Today, this number has ballooned to 8,168 species, a 456% increase (**http://www.bacterio.cict.fr/number.html#total**). The explosion in the number of recognized taxa is directly attributable to the ease in performance of 16S rRNA gene sequencing studies as opposed to the more cumbersome manipulations involving DNA-DNA hybridization investigations. DNA-DNA

Exercise 2

- RNAmmer Creates a HMM from the structural alignments to predict rRNA genes



SUBMISSION

Paste a single sequence or several sequences in FASTA format into the field below:
Select kingdom of input sequences:
Bacteria

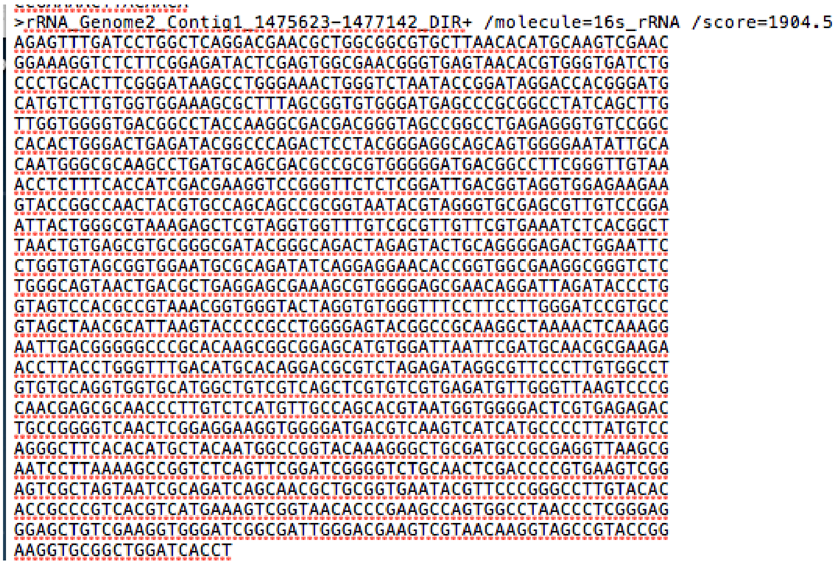
Submit a file in FASTA format directly from your local disk:
Choose File no file selected

Restrictions:
At most 10,000 sequences and 20,000,000 nucleotides per submission

Confidentiality:
The sequences are kept confidential and will be deleted after processing.

Exercise 2 cont’d

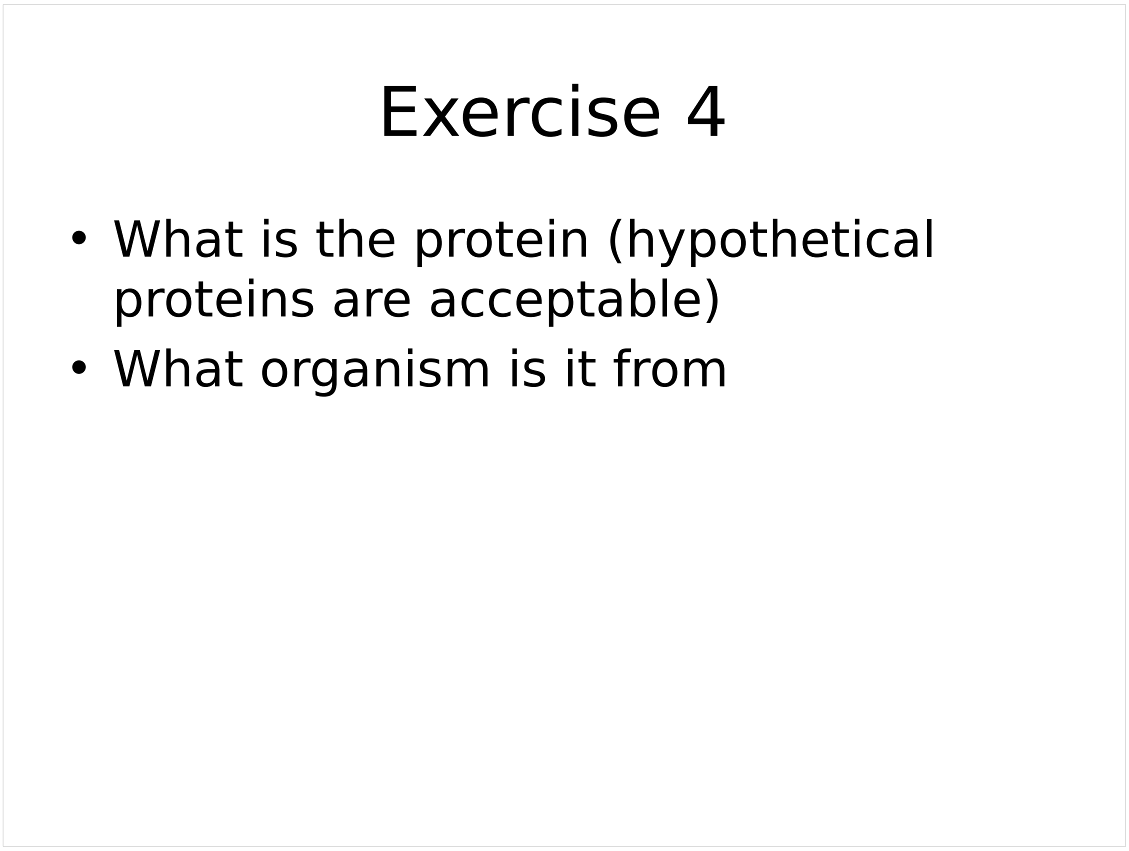
- Green genes Compare BLAST



```
>rRNA_Genome2_Contig1_1475623-1477142_DIR+ /molecule=16s_rRNA /score=1904.5
AGAGTTTGATCCTGGCTCAGGACGAACGCTGGCGGCGTGCTTAACACATGCAAGTCGAAC
GGAAAGGTCTCTTCGGAGATACTCGAGTGGCGAACGGGTGAGTAACACGTGGGTGATCTG
CCCTGCACTTCGGGATAAGCCTGGGAAACTGGGTCTAATACCGGATAGGACCACGGGATG
CATGTCTTGTGGTGGAAAGCGCTTTAGCGGTGTGGGATGAGCCCGCGGCCTATCAGCTTG
TTGGTGGGGTGACGGCCTACCAAGGCGACGACGGGTAGCCGGCCTGAGAGGGTGTCCGGC
CACACTGGGACTGAGATACGGCCCAGACTCCTACGGGAGGCAGCAGTGGGGAATATTGCA
CAATGGGCGCAAGCCTGATGCAGCGACGCCGCGTGGGGGATGACGGCCTTCGGGTTGTAA
ACCTCTTTCACCATCGACGAAGGTCCGGGTTCTCTCGGATTGACGGTAGGTGGAGAAGAA
GTACCGGCCAACTACGTGCCAGCAGCCGCGGTAATACGTAGGGTGCGAGCGTTGTCCGGA
ATTACTGGGCGTAAAGAGCTCGTAGGTGGTTTGTCGCGTTGTTCGTGAAATCTCACGGCT
TAACTGTGAGCGTGCGGGCGATACGGGCAGACTAGAGTACTGCAGGGGAGACTGGAATTC
CTGGTGTAGCGGTGGAATGCGCAGATATCAGGAGGAACACCGGTGGCGAAGGCGGGTCTC
TGGGCAGTAACTGACGCTGAGGAGCGAAAGCGTGGGGAGCGAACAGGATTAGATACCCTG
GTAGTCCACGCCGTAAACGGTGGGTACTAGGTGTGGGTTTCCTTCCTTGGGATCCGTGCC
GTAGCTAACGCATTAAGTACCCCGCCTGGGGAGTACGGCCGCAAGGCTAAAACTCAAAGG
AATTGACGGGGGCCCGCACAAGCGGCGGAGCATGTGGATTAATTCGATGCAACGCGAAGA
ACCTTACCTGGGTTTGACATGCACAGGACGCGTCTAGAGATAGGCGTTCCCTTGTGGCCT
GTGTGCAGGTGGTGCATGGCTGTCGTCAGCTCGTGTCGTGAGATGTTGGGTTAAGTCCCG
CAACGAGCGCAACCCTTGTCTCATGTTGCCAGCACGTAATGGTGGGGACTCGTGAGAGAC
TGCCGGGGTCAACTCGGAGGAAGGTGGGGATGACGTCAAGTCATCATGCCCCTTATGTCC
AGGGCTTCACACATGCTACAATGGCCGGTACAAAGGGCTGCGATGCCGCGAGGTTAAGCG
AATCCTTAAAAGCCGGTCTCAGTTCGGATCGGGGTCTGCAACTCGACCCCGTGAAGTCGG
AGTCGCTAGTAATCGCAGATCAGCAACGCTGCGGTGAATACGTTCCCGGGCCTTGTACAC
ACCGCCCGTCACGTCATGAAAGTCGGTAACACCCGAAGCCAGTGGCCTAACCCTCGGGAG
GGAGCTGTCGAAGGTGGGATCGGCGATTGGGACGAAGTCGTAACAAGGTAGCCGTACCGG
AAGGTGCGGCTGGATCACCT
```

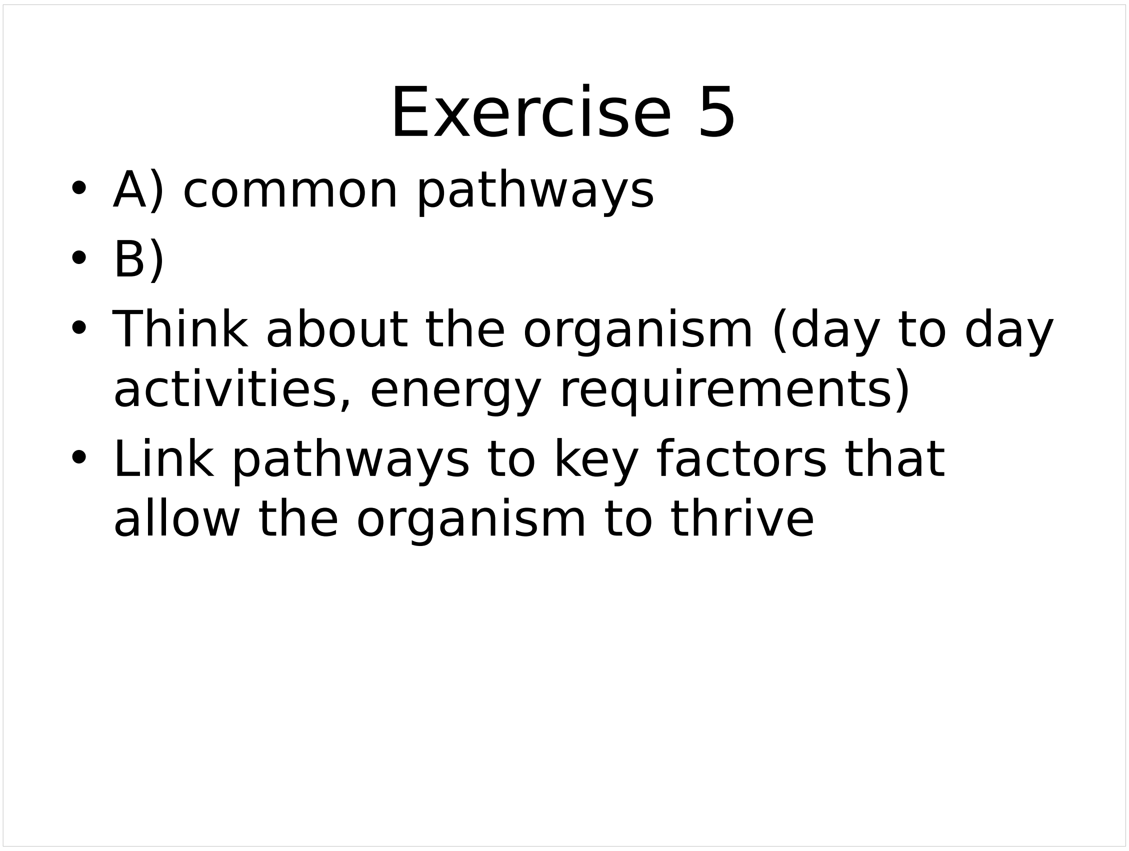
Exercise 3

- The reading frame
- Why? (think about how we select the best reading frame)



Exercise 4

- What is the protein (hypothetical proteins are acceptable)
- What organism is it from



Exercise 5

- A) common pathways
- B)
- Think about the organism (day to day activities, energy requirements)
- Link pathways to key factors that allow the organism to thrive

Exercise 6

- Look up each organism
- Noting differences
- Consider the pathways that are *different* between the organisms;
- Link the pathways to the differences
- Use **references** to scientific journals