Jacek Radajewski
Student number: 43612772
SCIE2100 Prac 7

Question 1

Results

Genome	Mean	STD DEV	Upper Bound	Lower Bound
genome1	0.38	0.09	0.56	0.2
genome2	0.64	0.02	0.68	0.6
genome3	0.39	0.04	0.47	0.31

Code

```
'''
Created on 13/05/2014

@author: jacekrad
'''
from sequence import *

fasta_files = ["Genome1.fasta", "Genome2.fasta", "Genome3.fasta"]

def get_gc_count(sequence):
    return sequence.count('G') + sequence.count('C')

def get_gc_fraction(sequence):
    return round(float(get_gc_count(sequence)) / len(sequence), 2)

def get_mean(values):
    return round(float(sum(values)) / len(values), 2)

def get_standard_deviation(values):
    vals = []
    mean = get_mean(values)
    for i in range(len(values)):
        vals.append((values[i] - mean) ** 2)
    standard_deviation = math.sqrt((1 / float(len(values))) * sum(vals))
    return round(standard_deviation, 2)

#dictionary in which we'll save the contigs
contigs = {}

for fasta_file in fasta_files:
    contig_list = []
    contigs.update({fasta_file:contig_list})
    sequences = readFastaFile(fasta_file, DNA_Alphabet)
    for sequence in sequences:
        contig_list.append(get_gc_fraction(sequence.sequence))
    mean = get_mean(contig_list)
    standard_deviation = get_standard_deviation(contig_list)
    upper_bound = mean + (2 * standard_deviation)
    lower_bound = mean - (2 * standard_deviation)
    print fasta_file, ": ", mean, standard_deviation, upper_bound,
lower_bound
```

Unit Test

```
import unittest
import question1 as q1

class Q1Test(unittest.TestCase):

    data1 = [1, 2, 3, 4, 5]
    data2 = [1, 2, 3, 4, 5, 6]

    def setUp(self):
        pass

    def tearDown(self):
        pass

    def test__mean__1(self):
        self.assertEquals(3, q1.get_mean(self.data1), "")

    def test__mean__2(self):
        self.assertEquals(3.5, q1.get_mean(self.data2), "")

    def test__standard_deviation__1(self):
        self.assertEquals(1.41, q1.get_standard_deviation(self.data1), "")

    def test__standard_deviation__2(self):
        self.assertEquals(1.71, q1.get_standard_deviation(self.data2), "")

if __name__ == "__main__":
    # import sys;sys.argv = ['', 'Test.testName']
    unittest.main()
```

Output

```
Genome1.fasta :  0.38 0.09 0.56 0.2
Genome2.fasta :  0.64 0.02 0.68 0.6
Genome3.fasta :  0.39 0.04 0.47 0.31
```

Question 2

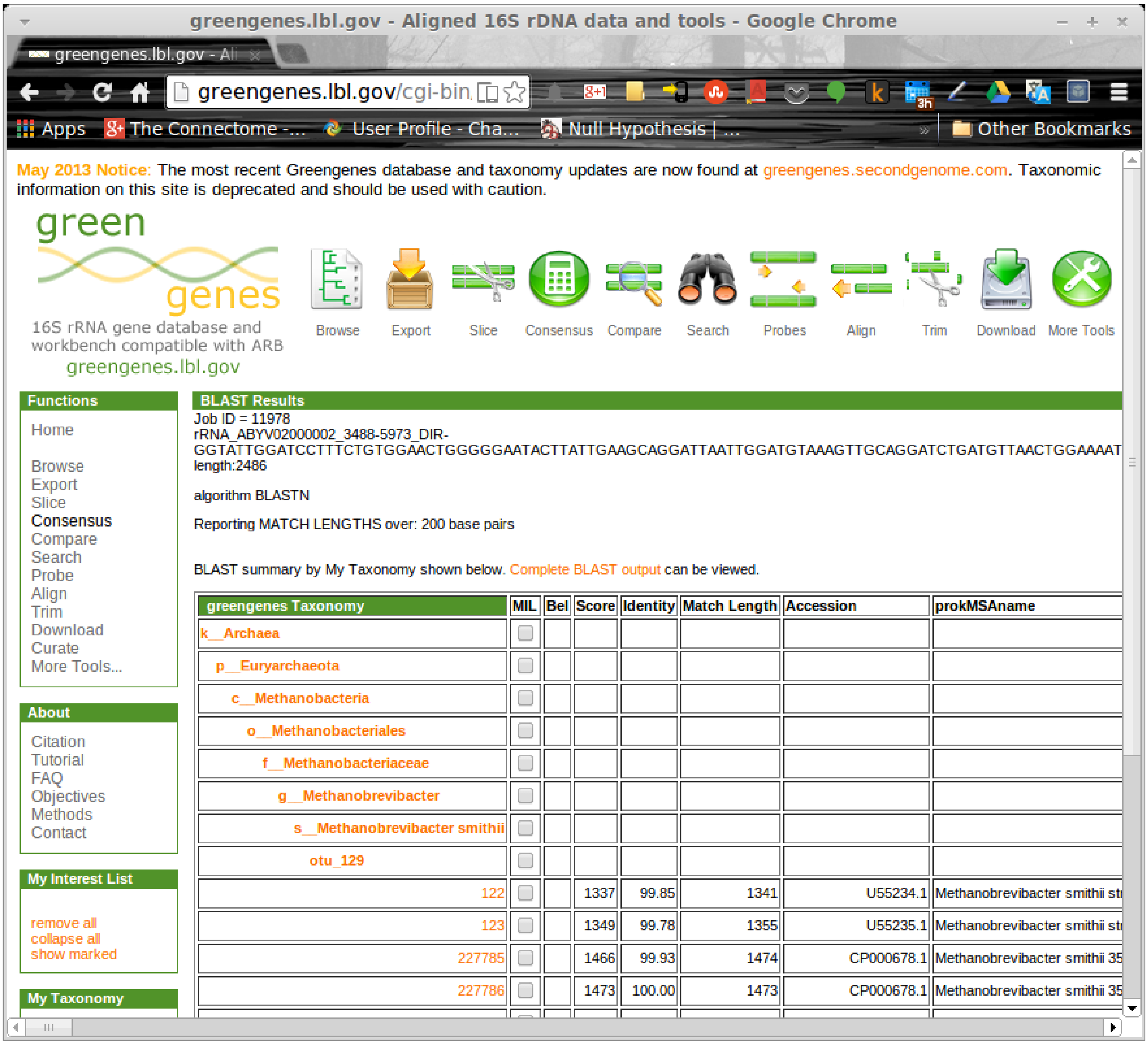
	Genome1	Genome2	Genome3
Kingdom	Archaea	Bacteria	Bacteria
Phylum	Euryarchaeota	Actinobacteria	Firmicutes
Class	Methanobacteria	Actinobacteria (class)	Clostridia
Order	Methanobacteriales	Actinomycetales	Clostridiales
Family	Methanobacteriaceae	Mycobacteriaceae	Veillonellaceae
Genus	Methanobrevibacter	Mycobacterium	Veillonella
Species	s__Methanobrevibacter smithii	Unclassified	Veillonella dispar

Genome1

16S rRNA

```
>rRNA_ABYV02000002_3488-5973_DIR- /molecule=16s_rRNA /score=612.8
GGTATTGGATCCTTTCTGTGGAACTGGGGGAATACTTATTGAAGCAGGATTAATTGGATG
TAAAGTTGCAGGATCTGATGTTAACTGGAAAATGAAAAATGGAAGTGCAATCAATTTGGA
TTACTGTGGAATAACTGATTATAGAACATTCAATGTTGATGTACGTGAACTTAAAATGTA
TGAAAAGGTAGATAGTGTAGTTACAGACCCTCCTTATGGAATATCTACTTCAACTGGGGA
TATTGAGGGTGATGAGATTTTCAATGAGTTTTTCCATTCAATTTATGATAATATGAAAGA
TGATGCCTACTTGTGTATGGCTAGTCCTCATTATGTTGATTTAAATCCTATGATTAAGGA
AGTTGGATTTGAATTAGTTGAACAATATGGAATCAAAATGCATAGAAGTTTAACAAGGAT
AATTTCAGTTATTCGTAAGAAAAATGTTTAATTTTTTTTATTTATTTATAAATAAATAGG
TAAGTTTATATATTACTTAGTAATCGATTTTACTTATTTATATATTATATTTGCTTTTTT
TCCAATTTATTTATTTTAAAAGAGATTTTTTACTTTTGCTTTTCATTTTGACTTTTAAAA
TTTTCATTAGGTCTATTTGATTAAATTTTTCATTTTTATGAATCATTAGTTTAACTATTG
TTTTTTTATAAATTAAAGATGTTTCATTAATATTTTTTTATTTGATTTATTTTTGCAATA
ATTTTTAATTGAATTTGTTGATATAATTTCCTTGTCTATTAGATGTGCTATGTGTTATAT
GGCGTTGGTCTAAGTTACATTGTATTGACAATTATAACTATGATGACTGCTTTTCCACAT
GTAGCTAAACATATTTAATGAGCACTTGAGTTTTTTGAGTGTGATGTTGGTTTTGTAGAT
GTGGTGAATTTGATTACATTATTGTTATCAAATCAGTGTATTTACTCGTCTATATTTTTT
TAGCGTACTTCATAAATTTTAGCTTTTTTATGATTATCTGCTTTTTTCATTCAATTCTGT
TTGATCCTGGCAGATGCTACTGCTATTGGGATTCGATTAAGCCATGCAAGTCGAACGAGT
TTAGGCTCGTGGCGTACGGCTCAGTAACACGTGGATAACCTACCCTTAGGACTGGGATAA
CCCTGGGAAACTGGGGATAATACTGGATAGGCAATTATTCCTGTAATGGTTTTTTGTTTA
AATGTTTTTTCGCCTAAGGATGGGTCTGCGGCCGATTAGGTAGTTGGTTAGGTAATGGCT
TACCAAGCCTTTGATCGGTACGGGTTGTGAGAGCAAGAGCCCGGAGATGGAACCTGAGAC
AAGGTTCCAGGCCCTACGGGGTGCAGCAGGCGCGAAACCTCCGCAATGTGAGAAATCGCG
ACGGGGGGATCCCAAGTGCCATTCTTAACGGGATGGCTTTTCATTAGTGTAAAGAGCTTT
TGGAATAAGAGCTGGGCAAGACCGGTGCCAGCCGCCGCGGTAACACCGGCAGCTCTAGTG
GTAGCAGTTTTTATTGGGCCTAAAGCGTCCGTAGCCGGTTTAATAAGTCTCTGGTGAAAT
CCTGCAGCTTAACTGTGGGAATTGCTGGAGATACTATTAGACTTGAGATCGGGAGAGGTT
AGAGGTACTCCCAGGGTAGAGGTGAAATTCTGTAATCCTGGGAGGACCGCCTGTTGCGAA
GGCGTCTGACTGGAACGATTCTGACGGTGAGGGACGAAAGCTAGGGGCGCGAACCGGATT
AGATACCCGGGTAGTCCTAGCTGTAAACGATGCGGACTTGGTGTTGGGGTGGCTTTGAGC
TGTCCCAGTGCCGAAGGGAAGCTGTTAAGTCCGCCGCCTGGGAAGTACGGTCGCAAGACT
GAAACTTAAAGGAATTGGCGGGGGAGCACCACAACGCGTGGAGCCTGCGGTTTAATTGGA
TTCAACGCCGGACATCTCACCAGAGGCGACAGCTGTATGATAGCCAGGTTGATGACTTTG
CTTGACTAGCTGAGAGGAGGTGCATGGCCGCCGTCAGCTCGTACCGTGAGGCGTCCTGTT
AAGTCAGGCAACGAGCGAGACCCACGCTCTTAGTTACCAGCGGATCCTTTTTTGGATGCC
GGGCACACTAAGGGGACCGCCTATGATAAATAGGAGGAAGGAGTGGACGACGGTAGGTCC
GTATGCCCCGAATCCTCTGGGCAACACGCGGGCTACAATGGCTGAGACAATGGGTTCCGA
CGCCGAAAGGCGGAGGTAATCCTCTAAACTTAGTCGTAGTTCGGATTGAGGACTGTAACT
CGTTCTCATGAAGCTGGAATGCGTAGTAATCGCGTGTCACAATCGCGCGGTGAATACGTC
CCTGCTCCTTGCACACACCGCCCGTCACGCCACCCAAAAAGGGATTGGATGAGGATGTAA
TGTTTTGTTATATTCGAATCTAGTTTTTTTAAGGAGGGCGAAGTCGTAACAAGGTAGCCG
TAGGGGAACCTGCGGCTGGATCACCT
```

Results



May 2013 Notice: The most recent Greengenes database and taxonomy updates are now found at greengenes.secondgenome.com. Taxonomic information on this site is deprecated and should be used with caution.

BLAST Results

Job ID = 11978
rRNA_ABYV02000002_3488-5973_DIR-
GGTATTGGATCCTTTCTGTGGAACTGGGGGAATACTTATTGAAGCAGGATTAATTGGATGTAAAGTTGCAGGATCTGATGTTAACTGGAAAAT
length:2486

algorithm BLASTN

Reporting MATCH LENGTHS over: 200 base pairs

BLAST summary by My Taxonomy shown below. Complete BLAST output can be viewed.

greengenes Taxonomy	MIL	Bel	Score	Identity	Match Length	Accession	prokMSAname
k__Archaea	☐						
p__Euryarchaeota	☐						
c__Methanobacteria	☐						
o__Methanobacteriales	☐						
f__Methanobacteriaceae	☐						
g__Methanobrevibacter	☐						
s__Methanobrevibacter smithii	☐						
otu_129	☐						
122	☐		1337	99.85	1341	U55234.1	Methanobrevibacter smithii st
123	☐		1349	99.78	1355	U55235.1	Methanobrevibacter smithii st
227785	☐		1466	99.93	1474	CP000678.1	Methanobrevibacter smithii 35
227786	☐		1473	100.00	1473	CP000678.1	Methanobrevibacter smithii 35

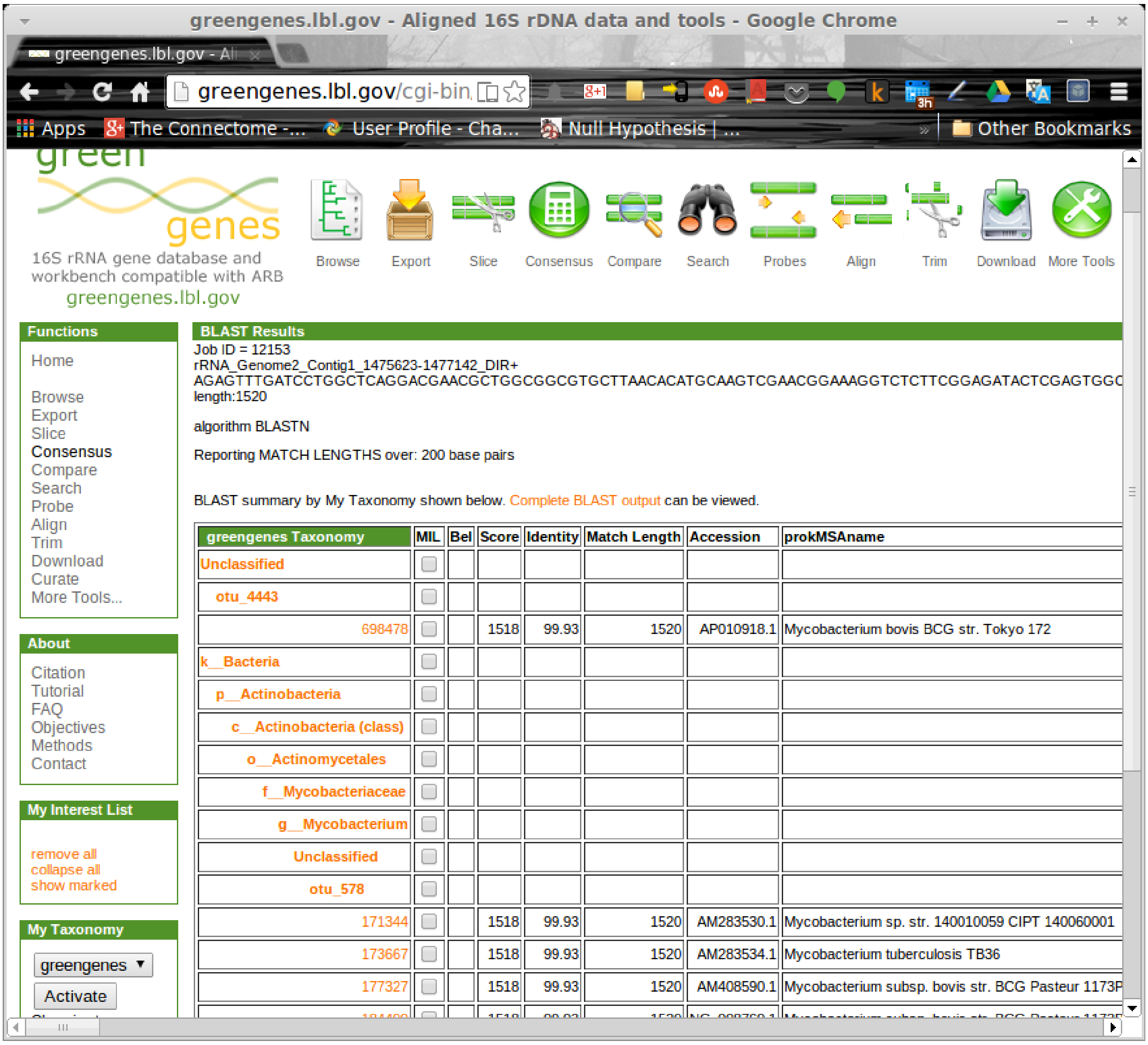
Genome2

16S rRNA

```
>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
```

Results



BLAST Results

Job ID = 12153
rRNA_Genome2_Contig1_1475623-1477142_DIR+
AGAGTTTGATCCTGGCTCAGGACGAACGCTGGCGGCGTGCTTAACACATGCAAGTCGAACGGAAAGGTCTCTTCGGAGATACTCGAGTGGC
length:1520

algorithm BLASTN

Reporting MATCH LENGTHS over: 200 base pairs

BLAST summary by My Taxonomy shown below. Complete BLAST output can be viewed.

greengenes Taxonomy	MIL	Bel	Score	Identity	Match Length	Accession	prokMSAname
Unclassified	□						
otu_4443	□						
698478	□		1518	99.93	1520	AP010918.1	Mycobacterium bovis BCG str. Tokyo 172
k__Bacteria	□						
p__Actinobacteria	□						
c__Actinobacteria (class)	□						
o__Actinomycetales	□						
f__Mycobacteriaceae	□						
g__Mycobacterium	□						
Unclassified	□						
otu_578	□						
171344	□		1518	99.93	1520	AM283530.1	Mycobacterium sp. str. 140010059 CIPT 140060001
173667	□		1518	99.93	1520	AM283534.1	Mycobacterium tuberculosis TB36
177327	□		1518	99.93	1520	AM408590.1	Mycobacterium subsp. bovis str. BCG Pasteur 1173P

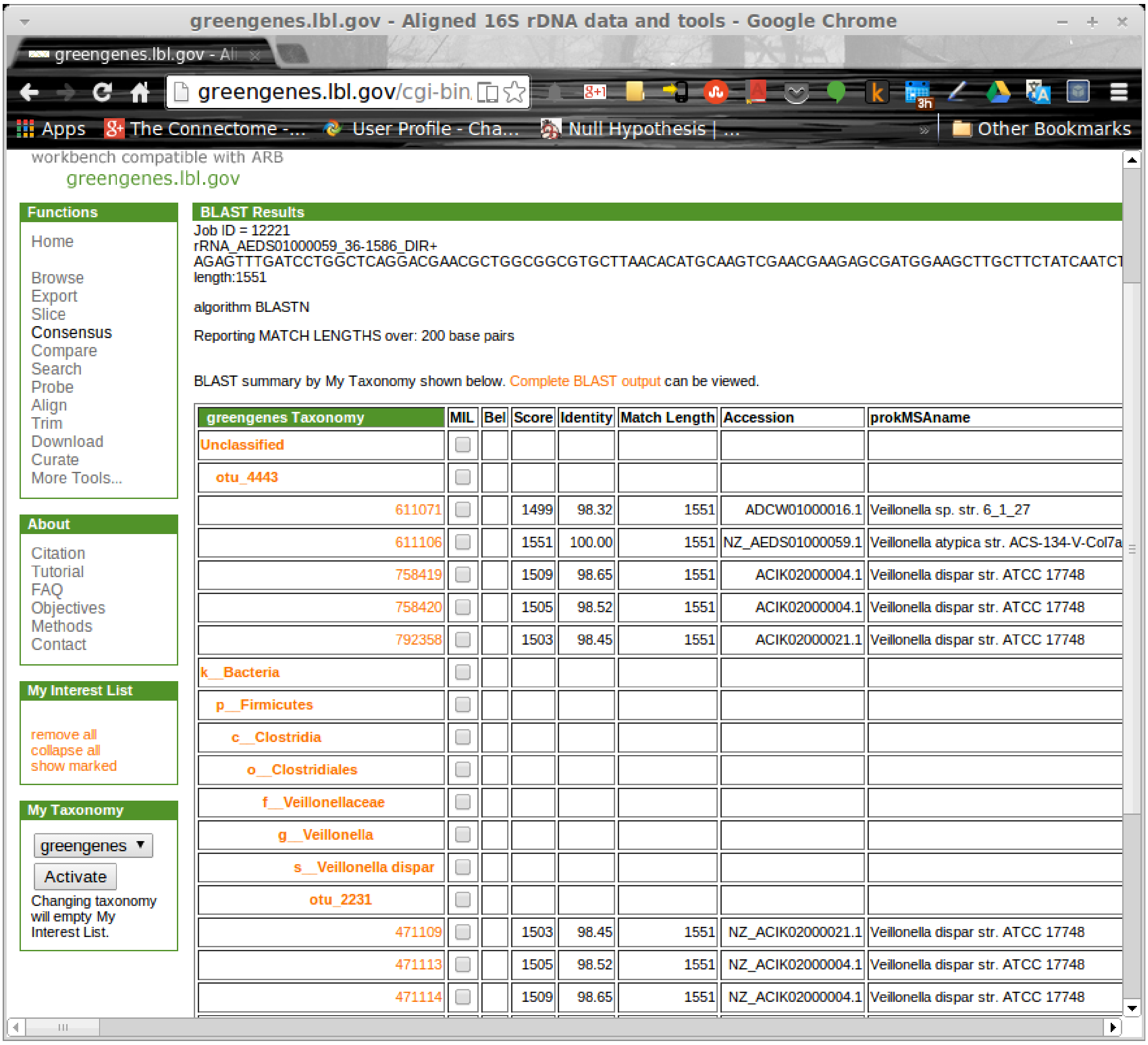
Genome3

16S rRNA

```
>rRNA_AEDS01000059_36-1586_DIR+ /molecule=16s_rRNA /score=1871.2
AGAGTTTGATCCTGGCTCAGGACGAACGCTGGCGGCGTGCTTAACACATGCAAGTCGAAC
GAAGAGCGATGGAAGCTTGCTTCTATCAATCTTAGTGGCGAACGGGTGAGTAACGCGTAA
TCAACCTGCCCTTCAGAGGGGGACAACAGTTGGAAACGACTGCTAATACCGCATACGATC
CAATCTCGGCATCGAGACTGGATGAAAGGTGGCCTCTATTTATAAGCTATCACTGAAGGA
GGGGATTGCGTCTGATTAGCTAGTTGGAGGGGTAACGGCCCACCAAGGCGATGATCAGTA
GCCGGTCTGAGAGGATGAACGGCCACATTGGGACTGAGACACGGCCCAGACTCCTACGGG
AGGCAGCAGTGGGGAATCTTCCGCAATGGACGAAAGTCTGACGGAGCAACGCCGCGTGAG
TGATGACGGCCTTCGGGTTGTAAAGCTCTGTTAATCGGGACGAATGGTTCTTGTGCGAAT
AGTGCGAGGATTTGACGGTACCGGAATAGAAAGCCACGGCTAACTACGTGCCAGCAGCCG
CGGTAATACGTAGGTGGCAAGCGTTGTCCGGAATTATTGGGCGTAAAGCGCGCGCAGGCG
GATTAGTTAGTCTGTCTTAAAAGTTCGGGGCTTAACCCCGTGATGGGATGGAAACTGCTG
ATCTAGAGTATCGGAGAGGAAAGTGGAATTCCTAGTGTAGCGGTGAAATGCGTAGATATT
AGGAAGAACACCAGTGGCGAAGGCGACTTTCTGGACGAAAACTGACGCTGAGGCGCGAAA
```

```
GCCAGGGGAGCGAACGGGATTAGATACCCCGGTAGTCCTGGCCGTAAACGATGGGTACTA
GGTGTAGGAGGTATCGACCCCTTCTGTGCCGGAGTTAACGCAATAAGTACCCCGCCTGGG
GAGTACGACCGCAAGGTTGAAACTCAAAGGAATTGACGGGGGCCCGCACAAGCGGTGGAG
TATGTGGTTTAATTCGACGCAACGCGAAGAACCTTACCAGGTCTTGACATTGATGGACAG
AACTAGAGATAGTTCCTCTTCTTCGGAAGCCAGAAAACAGGTGGTGCACGGTTGTCGTCA
GCTCGTGTCGTGAGATGTTGGGTTAAGTCCCGCAACGAGCGCAACCCCTATCTTATGTTG
CCAGCACTTCGGGTGGGAACTCATGAGAGACTGCCGCAGACAATGCGGAGGAAGGCGGGG
ATGACGTCAAATCATCATGCCCCTTATGACCTGGGCTACACACGTACTACAATGGGAGTT
AATAGACGGAAGCGAAACCGCGAGGTGGAGCAAACCCGAGAAACACTCTCTCAGTTCGGA
TCGTAGGCTGCAACTCGCCTACGTGAAGTCGGAATCGCTAGTAATCGCAGGTCAGCATAC
TGCGGTGAATACGTTCCCGGGCCTTGTACACACCGCCCGTCACACCACGAAAGTCGGAAG
TGCCCAAAGCCGGTGGGGTAACCTTCGGGAGCCAGCCGTCTAAGGTAAAGTCGATGATTG
GGGTGAAGTCGTAACAAGGTAGCCGTATCGGAAGGTGCGGCTGGATCACCT
```

Results



Changing taxonomy will empty My Interest List.

BLAST Results

Job ID = 12221
rRNA_AEDS01000059_36-1586_DIR+
AGAGTTTGATCCTGGCTCAGGACGAACGCTGGCGGCGTGCTTAACACATGCAAGTCGAACGAAGAGCGATGGAAGCTTGCTTCTATCAATCT
length:1551

algorithm BLASTN

Reporting MATCH LENGTHS over: 200 base pairs

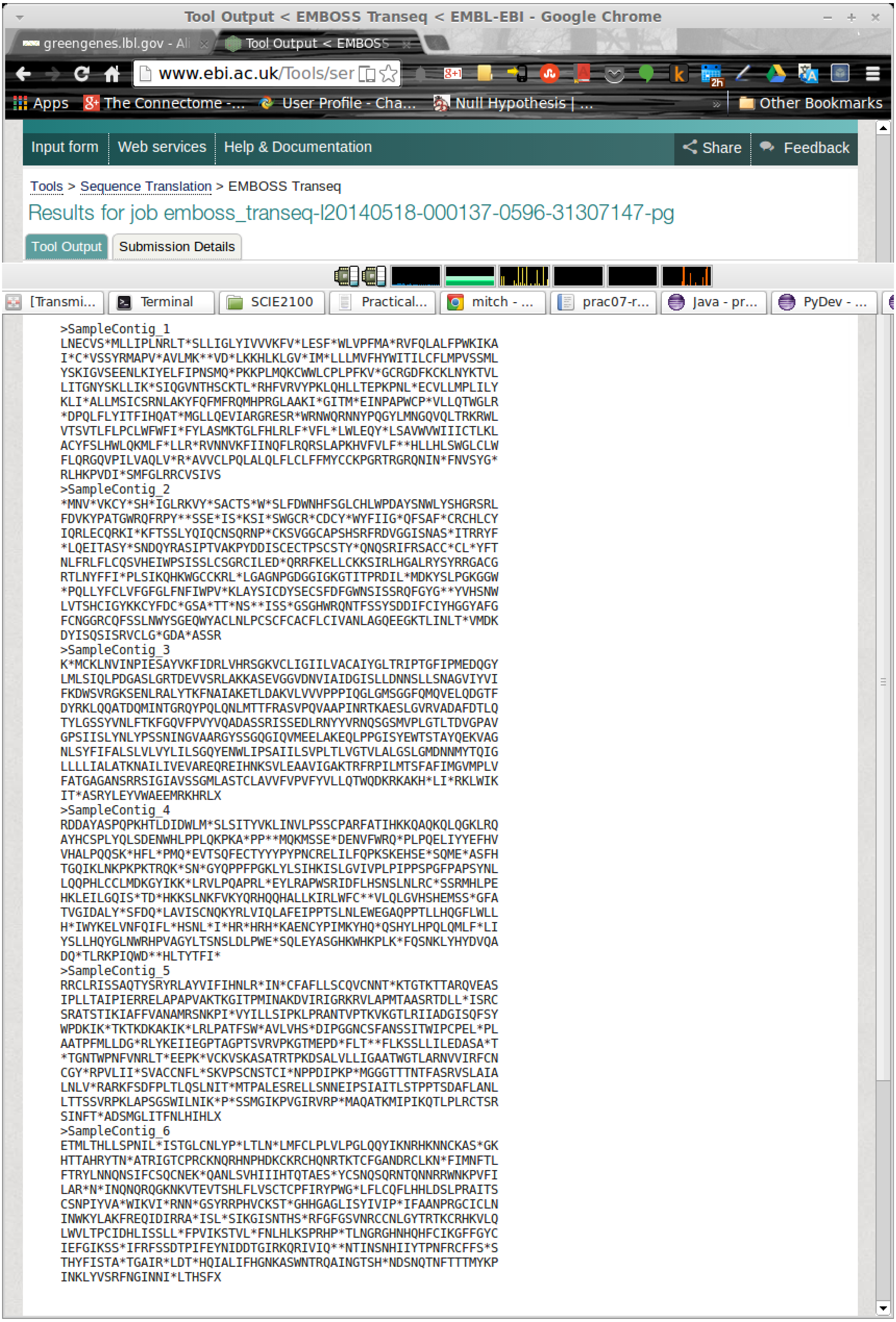
BLAST summary by My Taxonomy shown below. Complete BLAST output can be viewed.

greengenes Taxonomy	MIL	Bel	Score	Identity	Match Length	Accession	prokMSAname
Unclassified	□						
otu_4443	□						
611071	□		1499	98.32	1551	ADCW01000016.1	Veillonella sp. str. 6_1_27
611106	□		1551	100.00	1551	NZ_AEDS01000059.1	Veillonella atypica str. ACS-134-V-Col7a
758419	□		1509	98.65	1551	ACIK02000004.1	Veillonella dispar str. ATCC 17748
758420	□		1505	98.52	1551	ACIK02000004.1	Veillonella dispar str. ATCC 17748
792358	□		1503	98.45	1551	ACIK02000021.1	Veillonella dispar str. ATCC 17748
k__Bacteria	□						
p__Firmicutes	□						
c__Clostridia	□						
o__Clostridiales	□						
f__Veillonellaceae	□						
g__Veillonella	□						
s__Veillonella dispar	□						
otu_2231	□						
471109	□		1503	98.45	1551	NZ_ACIK02000021.1	Veillonella dispar str. ATCC 17748
471113	□		1505	98.52	1551	NZ_ACIK02000004.1	Veillonella dispar str. ATCC 17748
471114	□		1509	98.65	1551	NZ_ACIK02000004.1	Veillonella dispar str. ATCC 17748

Question 3

Out of the six reading frames (shown in screen grab) the following one is the correct translation because it starts with the correct START codon (M), ends with a STOP codon (*), and has a sufficient length to be a protein (over a hundred of residues).

>SampleContig_3
K***MCKLNVINPIESAYVKFIDRLVHRSGKVCLIGIILVACAIYGLTRIPTGFIPMEDQGY**
LMLSIQLPDGASLGRTDEVVSRLAKKASEVGGVDNVIAIDGISLLDNNSLLSNAGVIYVI
FKDWSVRGKSENLRALYTKFNAIAKETLDAKVLVVVPPPIQGLGMSGGFQMQVELQDGTF
DYRKLQQATDQMINTGRQYPQLQNLMTTFRASVPQVAAPINRTKAESLGVRVADAFDTLQ
TYLGSSYVNLFTKFGQVFPVYVQADASSRISSEDLRNYYVRNQSGSMVPLGTLTDVGPAV
GPSIISLYNLYPSSNINGVAARGYSSGQGIQVMEELAKEQLPPGISYEWTSTAYQEKVAG
NLSYFIFALSLVLVYLILSGQYENWLIPSAIILSVPLTLVGTVLALGSLGMDNNMYTQIG
LLLLIALATKNAILIVEVAREQREIHNKSVLEAAVIGAKTRFRPILMTSFAFIMGVMPLV
FATGAGANSRRSIGIAVSSGMLASTCLAVVFVPVFYVLLQTWQDKRKAKH*LI*RKLWIK
IT*ASRYLEYVWAEEMRKHRLX



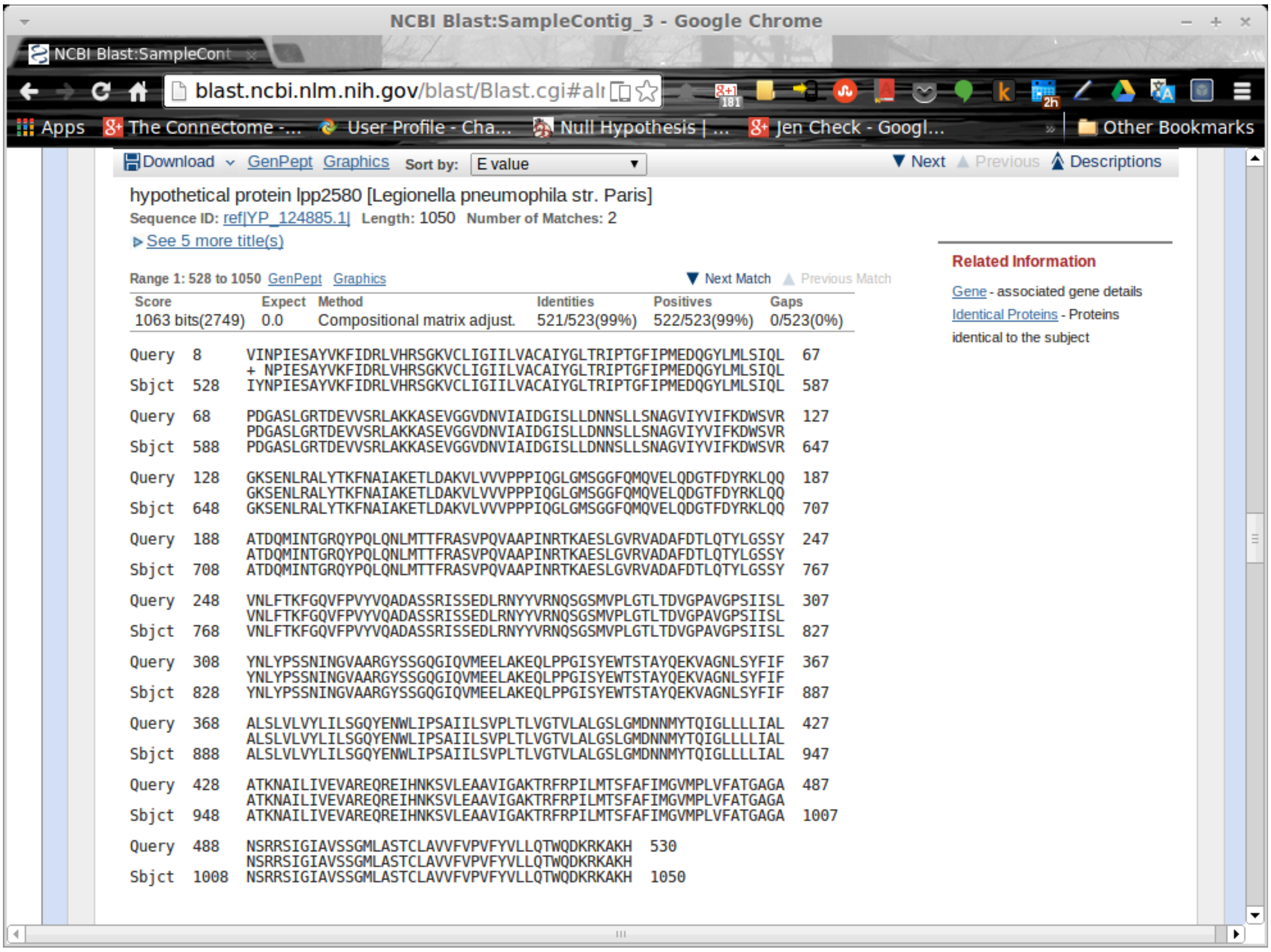
Tools > Sequence Translation > EMBOSS Transeq

Results for job emboss_transeq-I20140518-000137-0596-31307147-pg

```
>SampleContig_1
LNECVS*MLLIPLNRLT*SLLIGLYIVVVKFV*LESF*WLVPFMA*RVFQLALFPWKIKA
I*C*VSSYRMAPV*AVLMK**VD*LKKHLKLGV*IM*LLLMVFHYWITILCFLMPVSSML
YSKIGVSEENLKIYELFIPNSMQ*PKKPLMQKCWWLCPLPFKV*GCRGDFKCKLNYKTVL
LITGNYSKLLIK*SIQGVNTHSCKTL*RHFVRVYPKLQHLLTEPKPNL*ECVLLMPLILY
KLI*ALLMSICSRNLAKYFQFMFRQMHPRGLAAKI*GITM*EINPAPWCP*VLLQTWGLR
*DPQLFLYITFIHQAT*MGLLQEVIARGRESR*WRNWQRNNYPQGYLMNGQVQLTRKRWL
VTSVTLFLPCLWFWFI*FYLASMKTGLFHLRLF*VFL*LWLEQY*LSAVWVWIIICTLKL
ACYFSLHWLQKMLF*LLR*RVNNVKFIINQFLRQRSLAPKHVFVLF**HLLHLSWGLCLW
FLQRGQVPILVAQLV*R*AVVCLPQLALQLFLCLFFMYCCKPGRTRGRQNIN*FNVSYG*
RLHKPVDI*SMFGLRRCVSIVS
>SampleContig_2
*MNV*VKCY*SH*IGLRKVY*SACTS*W*SLFDWNHFSGLCHLWPDAYSNWLYSHGRSRL
FDVKYPATGWRQFRPY**SSE*IS*KSI*SWGCR*CDCY*WYFIIG*QFSAF*CRCHLCY
IQRLECQRKI*KFTSSLYQIQCNSQRNP*CKSVGGCAPSHSRFRDVGGISNAS*ITRRYF
*LQEITASY*SNDQYRASIPTVAKPYDDISCECTPSCSTY*QNQSRIFRSACC*CL*YFT
NLFRLFLCQSVHEIWPSISSLCSGRCILED*QRRFKELLCKKSIRLHGALRYSYRRGACG
RTLNYFFI*PLSIKQHKWGCCKRL*LGAGNPGDGGIGKGTITPRDIL*MDKYSLPGKGGW
*PQLLYFCLVFGFGLFNFIWPV*KLAYSICDYSECSFDFGWNSISSRQFGYG**YVHSNW
LVTSHCIGYKKCYFDC*GSA*TT*NS**ISS*GSGHWRQNTFSSYSDDIFCIYHGGYAFG
FCNGGRCQFSSLNWYSGEQWYACLNLPCSCFCACFLCIVANLAGQEEGKTLINLT*VMDK
DYISQSISRVCLG*GDA*ASSR
>SampleContig_3
K*MCKLNVINPIESAYVKFIDRLVHRSGKVCLIGIILVACAIYGLTRIPTGFIPMEDQGY
LMLSIQLPDGASLGRTDEVVSRLAKKASEVGGVDNVIAIDGISLLDNNSLLSNAGVIYVI
FKDWSVRGKSENLRALYTKFNAIAKETLDAKVLVVVPPPIQGLGMSGGFQMQVELQDGTF
DYRKLQQATDQMINTGRQYPQLQNLMTTFRASVPQVAAPINRTKAESLGVRVADAFDTLQ
TYLGSSYVNLFTKFGQVFPVYVQADASSRISSEDLRNYYVRNQSGSMVPLGTLTDVGPAV
GPSIISLYNLYPSSNINGVAARGYSSGQGIQVMEELAKEQLPPGISYEWTSTAYQEKVAG
NLSYFIFALSLVLVYLILSGQYENWLIPSAIILSVPLTLVGTVLALGSLGMDNNMYTQIG
LLLLIALATKNAILIVEVAREQREIHNKSVLEAAVIGAKTRFRPILMTSFAFIMGVMPLV
FATGAGANSRRSIGIAVSSGMLASTCLAVVFVPVFYVLLQTWQDKRKAKH*LI*RKLWIK
IT*ASRYLEYVWAEEMRKHRLX
>SampleContig_4
RDDAYASPQPKHTLDIDWLM*SLSITYVKLINVLPSSCPARFATIHKKQAQKQLQGKLRQ
AYHCSPLYQLSDENWHLPPLQKPKA*PP**MQKMSSE*DENVFWRQ*PLPQELIYYEFHV
VHALPQQSK*HFL*PMQ*EVTSQFECTYYYPYPNCRELILFQPKSKEHSE*SQME*ASFH
TGQIKLNKPKPKTRQK*SN*GYQPPFPGKLYLSIHKISLGVIVPLPIPPSPGFPAPSYNL
LQQPHLCCLMDKGYIKK*LRVLPQAPRL*EYLRAPWSRIDFLHSNSLNLRC*SSRMHLPE
HKLEILGQIS*TD*HKKSLNKFVKYQRHQQHALLKIRLWFC**VLQLGVHSHEMSS*GFA
TVGIDALY*SFDQ*LAVISCNQKYRLVIQLAFEIPPTSLNLEWEGAQPPTLLHQGFLWLL
H*IWYKELVNFQIFL*HSNL*I*HR*HRH*KAENCYPIMKYHQ*QSHYLHPQLQMLF*LI
YSLLHQYGLNWRHPVAGYLTSNSLDLPWE*SQLEYASGHKWHKPLK*FQSNKLYHYDVQA
DQ*TLRKPIQWD**HLTYTFI*
>SampleContig_5
RRCLRISSAQTYSRYRLAYVIFIHNLR*IN*CFAFLLSCQVCNNT*KTGTKTTARQVEAS
IPLLTAIPIERRELAPAPVAKTKGITPMINAKDVIRIGRKRVLAPMTAASRTDLL*ISRC
SRATSTIKIAFFVANAMRSNKPI*VYILLSIPKLPRANTVPTKVKGTLRIIADGISQFSY
WPDKIK*TKTKDKAKIK*LRLPATFSW*AVLVHS*DIPGGNCSFANSSITWIPCPEL*PL
AATPFMLLDG*RLYKEIIEGPTAGPTSVRVPKGTMEPD*FLT**FLKSSLLILEDASA*T
*TGNTWPNFVNRLT*EEPK*VCKVSKASATRTPKDSALVLLIGAATWGTLARNVVIRFCN
CGY*RPVLII*SVACCNFL*SKVPSCNSTCI*NPPDIPKP*MGGGTTTNTFASRVSLAIA
LNLV*RARKFSDFPLTLQSLNIT*MTPALESRELLSNNEIPSIAITLSTPPTSDAFLANL
LTTSSVRPKLAPSGSWILNIK*P*SSMGIKPVGIRVRP*MAQATKMIPIKQTLPLRCTSR
SINFT*ADSMGLITFNLHIHLX
>SampleContig_6
ETMLTHLLSPNIL*ISTGLCNLYP*LTLN*LMFCLPLVLPGLQQYIKNRHKNNCKAS*GK
HTTAHRYTN*ATRIGTCPRCKNQRHNPHDKCKRCHQNRTKTCFGANDRCLKN*FIMNFTL
FTRYLNNQNSIFCSQCNEK*QANLSVHIIIHTQTAES*YCSNQSQRNTQNNRRWNKPVFI
LAR*N*INQNQRQGKNKVTEVTSHLFLVSCTCPFIRYPWG*LFLCQFLHHLDSLPRAITS
CSNPIYVA*WIKVI*RNN*GSYRRPHVCKST*GHHGAGLISYIVIP*IFAANPRGCICLN
INWKYLAKFREQIDIRRA*ISL*SIKGISNTHS*RFGFGSVNRCCNLGYTRTKCRHKVLQ
LWVLTPCIDHLISSLL*FPVIKSTVL*FNLHLKSPRHP*TLNGRGHNHQHFCIKGFFGYC
IEFGIKSS*IFRFSSDTPIFEYNIDDTGIRKQRIVIQ**NTINSNHIIYTPNFRCFFS*S
THYFISTA*TGAIR*LDT*HQIALIFHGNKASWNTRQAINGTSH*NDSNQTNFTTTMYKP
INKLYVSRFNGINNI*LTHSFX
```

Question 4

The first result in the list was a hypothetical protein lpp2580 (shown in screen grab below). Uniprot search for this protein tells us that it is from **Legionella pneumophila (strain Paris)** species and it's taxonomic lineage is: Bacteria > Proteobacteria > Gammaproteobacteria > Legionellales › Legionellaceae › Legionella



Sort by: E value

hypothetical protein lpp2580 [Legionella pneumophila str. Paris]

Sequence ID: ref|YP_124885.1| Length: 1050 Number of Matches: 2

▸ See 5 more title(s)

Range 1: 528 to 1050 GenPept Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps
1063 bits(2749)	0.0	Compositional matrix adjust.	521/523(99%)	522/523(99%)	0/523(0%)

```
Query  8     VINPIESAYVKFIDRLVHRSGKVCLIGIILVACAIYGLTRIPTGFIPMEDQGYLMLSIQL  67
             + NPIESAYVKFIDRLVHRSGKVCLIGIILVACAIYGLTRIPTGFIPMEDQGYLMLSIQL
Sbjct  528   IYNPIESAYVKFIDRLVHRSGKVCLIGIILVACAIYGLTRIPTGFIPMEDQGYLMLSIQL  587

Query  68    PDGASLGRTDEVVSRLAKKASEVGGVDNVIAIDGISLLDNNSLLSNAGVIYVIFKDWSVR  127
             PDGASLGRTDEVVSRLAKKASEVGGVDNVIAIDGISLLDNNSLLSNAGVIYVIFKDWSVR
Sbjct  588   PDGASLGRTDEVVSRLAKKASEVGGVDNVIAIDGISLLDNNSLLSNAGVIYVIFKDWSVR  647

Query  128   GKSENLRALYTKFNAIAKETLDAKVLVVVPPPIQGLGMSGGFQMQVELQDGTFDYRKLQQ  187
             GKSENLRALYTKFNAIAKETLDAKVLVVVPPPIQGLGMSGGFQMQVELQDGTFDYRKLQQ
Sbjct  648   GKSENLRALYTKFNAIAKETLDAKVLVVVPPPIQGLGMSGGFQMQVELQDGTFDYRKLQQ  707

Query  188   ATDQMINTGRQYPQLQNLMTTFRASVPQVAAPINRTKAESLGVRVADAFDTLQTYLGSSY  247
             ATDQMINTGRQYPQLQNLMTTFRASVPQVAAPINRTKAESLGVRVADAFDTLQTYLGSSY
Sbjct  708   ATDQMINTGRQYPQLQNLMTTFRASVPQVAAPINRTKAESLGVRVADAFDTLQTYLGSSY  767

Query  248   VNLFTKFGQVFPVYVQADASSRISSEDLRNYYVRNQSGSMVPLGTLTDVGPAVGPSIISL  307
             VNLFTKFGQVFPVYVQADASSRISSEDLRNYYVRNQSGSMVPLGTLTDVGPAVGPSIISL
Sbjct  768   VNLFTKFGQVFPVYVQADASSRISSEDLRNYYVRNQSGSMVPLGTLTDVGPAVGPSIISL  827

Query  308   YNLYPSSNINGVAARGYSSGQGIQVMEELAKEQLPPGISYEWTSTAYQEKVAGNLSYFIF  367
             YNLYPSSNINGVAARGYSSGQGIQVMEELAKEQLPPGISYEWTSTAYQEKVAGNLSYFIF
Sbjct  828   YNLYPSSNINGVAARGYSSGQGIQVMEELAKEQLPPGISYEWTSTAYQEKVAGNLSYFIF  887

Query  368   ALSLVLVYLILSGQYENWLIPSAIILSVPLTLVGTVLALGSLGMDNNMYTQIGLLLLIAL  427
             ALSLVLVYLILSGQYENWLIPSAIILSVPLTLVGTVLALGSLGMDNNMYTQIGLLLLIAL
Sbjct  888   ALSLVLVYLILSGQYENWLIPSAIILSVPLTLVGTVLALGSLGMDNNMYTQIGLLLLIAL  947

Query  428   ATKNAILIVEVAREQREIHNKSVLEAAVIGAKTRFRPILMTSFAFIMGVMPLVFATGAGA  487
             ATKNAILIVEVAREQREIHNKSVLEAAVIGAKTRFRPILMTSFAFIMGVMPLVFATGAGA
Sbjct  948   ATKNAILIVEVAREQREIHNKSVLEAAVIGAKTRFRPILMTSFAFIMGVMPLVFATGAGA  1007

Query  488   NSRRSIGIAVSSGMLASTCLAVVFVPVFYVLLQTWQDKRKAKH  530
             NSRRSIGIAVSSGMLASTCLAVVFVPVFYVLLQTWQDKRKAKH
Sbjct  1008  NSRRSIGIAVSSGMLASTCLAVVFVPVFYVLLQTWQDKRKAKH  1050
```

Related Information

Gene - associated gene details

Identical Proteins - Proteins identical to the subject

Question 5

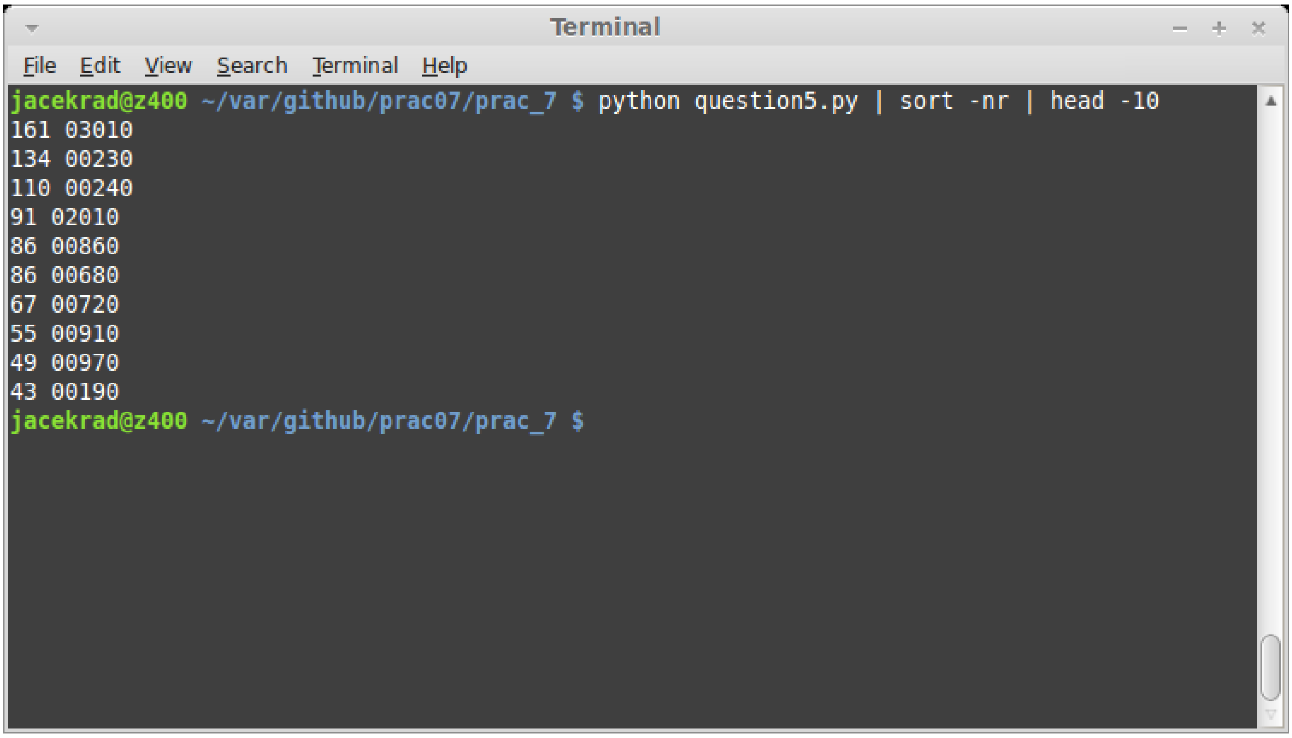
Top 10 Pathways were calculated based on the counts (scores) provided. For each pathway the scores from each genome were added to give the final score (see code and screenshot below). Pathways with top 10 scores/counts, from highest to lowest were:

Code	Score	Name
03010	161	Ribosome
00230	134	Purine metabolism
00240	110	Pyrimidine metabolism
02010	91	ABC Transporters
00860	86	Porphyrin and chlorophyll metabolism
00680	86	Methane metabolism
00720	67	Carbon fixation pathways in prokaryotes
00910	55	Nitrogen metabolism
00970	49	Aminoacyl-tRNA biosynthesis
00190	43	Oxidative phosphorylation

Note that the above list includes pathways with highest counts and some of these do not appear in all 3 genomes. Those pathways have been highlighted in cyan.

Above (non-highlighted) pathways are conserved because they are essential to basic function of the cell. For example:

- Ribosome: mRNA translation without which proteins could not be produced
- Purine and Pyrimidine metabolism is what creates nucleotides without which the cell would not be able to function



```
jacekrad@z400 ~/var/github/prac07/prac_7 $ python question5.py | sort -nr | head -10
161 03010
134 00230
110 00240
91 02010
86 00860
86 00680
67 00720
55 00910
49 00970
43 00190
jacekrad@z400 ~/var/github/prac07/prac_7 $
```

```
'''
Created on 19/05/2014

@author: jacekrad
'''


genome1 = {"00680":86, "03010":59, "00230":36, "00240":35, "00860":24, \
           "00970":24, "00720":19, "00400":18, "02010":18, "00250":16}
genome2 = {"03010":53, "00230":51, "00190":43, "02020":40, "00240":37, \
           "00910":34, "02010":33, "00860":30, "00720":27, "00330":27}
genome3 = {"03010":49, "00230":47, "02010":40, "00240":38, "00860":32, \
           "00970":25, "00270":24, "00720":21, "00910":21, "00400":19}

genomes = [genome1, genome2, genome3]

scores = {}

for genome in genomes:
    for pathway in genome:
        score = genome.get(pathway)
        existing = scores.get(pathway)
        if existing == None:
            scores.update({pathway:score})
        else:
            scores.update({pathway:(existing + score)})

for key in scores:
    print scores.get(key), key
```

Question 6

Methanotrophic archaeon

Genome 1 is the most likely candidate for this organism as it is the genome that shows methane metabolism.

Feng-Ping Wang, Yu Zhang, Ying Chen, Ying He, Ji Qi, Kai-Uwe Hinrichs, Xin-Xu Zhang, Xiang Xiao and Nico Boon, *Methanotrophic archaea possessing diverging methane-oxidizing and electron-transporting pathways*, The ISME Journal (2014) 8, 1069-1078

Veilonella

Genome3 is the most likeley match for this organism based on Phenylalanine, tyrosine and tryptophan biosynthesis.

http://patricbrc.org/portal/portal/patric/CompPathwayMap?cType=genome&cId=168093&dm=feature&feature_info_id=41230233&map=00970&algorithm=PATRIC&ec_number=

Mycobacterium

Genome2 is the most likely candidate for this organism based on the Arginine and proline metabolism.

Anjali Seth and Nancy D. Connell, *Amino Acid Transport and Metabolism in Mycobacteria: Cloning, Interruption, and Characterization of anL-Arginine/γ-Aminobutyric Acid Permease inMycobacterium bovis BCG,* February 2000, Journal of Bacteriology