BIOINFORMATICS

TA 2

==========

Name – Atharva Paliwal

Roll no. – B40

==========

1. Go to the main NCBI site: http://www.ncbi.nlm.nih.gov/ and search BS001137 and give the following information:

a.ACCESSION number: BS001137

b. DEFINITION: Severe acute respiratory syndrome coronavirus 2
hCov-19/Japan/KH879/2021 RNA, complete genome.

c. SOURCE: Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2)

d. REFERENCE: 1

e. FEATURES list: Location/Qualifiers

source 1..29863
/organism="Severe acute respiratory syndrome coronavirus 2"
/mol_type="genomic RNA"
/isolate="hCov-19/Japan/KH879/2021"
/isolation_source="saliva"
/host="Homo sapiens"
/db_xref="taxon:2697049"
/country="Japan: Tochigi"
/collection_date="2021-06-25"

5'UTR 1..243

gene 244..21524
/gene="ORF1ab"

f. CDs: join(244..13437,13437..21524)
/gene="ORF1ab"
/ribosomal_slippage

/codon_start=1

/product="ORF1ab polyprotein"

/protein_id="BDA76971.1"

/translation="MESLVPGFNEKTHVQLSLPVLQVRDVLVRGFGDSVEEVLSEARQ
HLKDGTCGLVEVEKGVLPQLEQPYVFIKRSDARTAPHGHVMVELVAELEGIQYGRSGE
TLGVLVPHVGEIPVAYRKVLLRKNGNKGAGGHSYGADLKSFDLGDELGTDPYEDFQEN
WNTKHSSGVTRELMRELNGGAYTRYVDNNFCGPDGYPLECIKDLLARAGKASCTLSEQ
LDFIDTKRGVYCCREHEHEIAWYTERSEKSYELQTPFEIKLAKKFDTFNGECPNFVFP
LNSIIKTIQPRVEKKKLDGFMGRIRSVYPVASPNECNQMCLSTLMKCDHCGETSWQTG
DFVKATCELCGTENLTKEGATTCGYLPQNAVVKIYCPACHNSEVGPEHSLAEYHNESG
LKTILRKGGRTIAFGGCVFSYVGCHNKCAYWVPRASANIGCNHTGVVGEGSEGLNDNL
LEILQKEKVNINIVVDFKLNEEIAIILASFSASTSAFVETVKGLDYKAFKQIVESCGN
FKVTKGKAKKGAWNIGEQKSILSPLYAFASEAARVVRSIFSRTLETAQNSVRVLQKAA
ITILDGISQYSLRLIDAMMFTSDLATNNLVVMAYITGGVVQLTSQWLTNIFGTVYEKL
KPVLDWLEEKFKEGVEFLRDGWEIVKFISTCACEIVGGQIVTCAKEIKESVQTFFKLV
NKFLALCADSIIIGGAKLKALNLGETFVTHSKGLYRKCVKSREETGLLMPLKAPKEII
FLEGETLPTEVLTEEVVLKTGDLQPLEQPTSEAVEAPLVGTPVCINGLMLLEIKDTEK
YCALAPNMMVTNNTFTLKGGAPTKVTFGDDTVIEVQGYKSVNITFELDERIDKVLNEK
CSAYTVELGTEVNEFACVVADAVIKTLQPVSELLTPLGIDLDEWSMATYYLFDESGEF
KLASHMYCSFYPPDEDEEEGDCEEEEFEPSTQYEYGTEDDYQGKPLEFGATSAALQPE
EEQEEDWLDDDSQQTVGQQDGSEDNQTTIIQTIVEVQPQLEMELTPVVQTIEVNSFSG
YLKLTDNVYIKNADIVEEAKKVKPTVVVNAANVYLKHGGGVAGALNKATNNAMQVESD
DYIATNGPLKVGGSCVLSGHNLAKHCLHVVGPNVNKGEDIQLLKSAYENFNQHEVLLA
PLLSAGIFGADPIHSLRVCVDTVRTNVYLAVFDKNLYDKLVSSFLEMKSEKQVEQKIA
EIPKEEVKPFITESKPSVEQRKQDDKKIKACVEEVTTTLEETKFLTENLLLYIDINGN
LHPDSATLVSDIDITFLKKDAPYIVGDVVQEGVLTAVVIPTKKAGGTTEMLAKALRKV
PTDNYITTYPGQGLNGYTVEEAKTVLKKCKSAFYILPSIISNEKQEILGTVSWNLREM
LAHAEETRKLMPVCVETKAIVSTIQRKYKGIKIQEGVVDYGARFYFYTSKTTVASLIN
TLNDLNETLVTMPLGYVTHGLNLEEAARYMRSLKVPATVSVSSPDAVTAYNGYLTSSS
KTPEEHFIETISLAGSYKDWSYSGQSTQLGIEFLKRGDKSVYYTSNPTTFHLDGEVIT
FDNLKTLLSLREVRTIKVFTTVDNINLHTQVVDMSMTYGQQFGPTYLDGADVTKIKPH
NSHEGKTFYVLPNDDTLRVEAFEYYHTTDPSFLGRYMSALNHTKKWKYPQVNGLTSIK

WADNNCYLATALLTLQQIELKFNPPALQDAYYRARAGEADNFCALILAYCNKTVGELG
DVRETMSYLFQHANLDSCKRVLNVVCKTCGQQQTTLKGVEAVMYMGTLSYEQFKKGVQ
IPCTCGKQATKYLVQQESPFVMMSAPPAQYELKHGTFTCASEYTGNYQCGHYKHITSK
ETLYCIDGALLTKSSEYKGPITDVFYKENSYTTTIKPVTYKLDGVVCTEIDPKLDNYY
KKDNSYFTEQPIDLVPNQPYPNASFDNFKFVCDNIKFADDLNQLTGYKKPASRELKVT
FFPDLNGDVVAIDYKHYTPSFKKGAKLLHKPIVWHVNNATNKATYKPNTWCIRCLWST
KPVETSNSFDVLKSEDAQGMDNLACEDLKPVSEEVVENPTIQKDVLECNVKTTEVVGD
IILKPANNSLKITEEVGHTDLMAAYVDNSSLTIKKPNELSRVLGLKTLATHGLAAVNS
VPWDTIANYAKPFLNKVVSTTTNIVTRCLNRVCTNYMPYFFTLLLQLCTFTRSTNSRI
KASMPTTIAKNTVKSVGKFCLEASFNYLKSPNFSKLINITIWFLLLSVCLGSLIYSTA
ALGVLMSNLGMPSYCTGYREGYLNSTNVTIATYCTGSIPCSVCLSGLDSLDTYPSLET
IQITISSFKWDLTAFGLVAEWFLAYILFTRFFYVLGLAAIMQLFFSYFAVHFISNSWL
MWLIINLVQMAPISAMVRMYIFFASFYYVWKSYVHVVDGCNSSTCMMCYKRNRATRVE
CTTIVNGVRRSFYVYANGGKGFCKLHNWNCVNCDTFCAGSTFISDEVARDLSLQFKRP
INPTDQSSYIVDSVTVKNGSIHLYFDKAGQKTYERHSLSHFVNLDNLRANNTKGSLPI
NVIVFDGKSKCEESSAKSASVYYSQLMCQPILLLDQALVSDVGDSAEVAVKMFDAYVN
TFSSTFNVPMEKLKTLVATAEAELAKNVSLDNVLSTFISAARQGFVDSDVETKDVVEC
LKLSHQSDIEVTGDSCNNYMLTYNKVENMTPRDLGACIDCSARHINAQVAKSHNIALI
WNVKDFMSLSEQLRKQIRSAAKKNNLPFKLTCATTRQVVNVVTTKIALKGGKIVNNWL
KQLIKVTLVFLFVAAIFYLITPVHVMSKHTDFSSEIIGYKAIDGGVTRDIASTDTCFA
NKHADFDTWFSQRGGSYTNDKACPLIAAVITREVGFVVPGLPGTILRTTNGDFLHFLP
RVFSAVGNICYTPSKLIEYTDFATSACVLAAECTIFKDASGKPVPYCYDTNVLEGSVA
YESLRPDTRYVLMDGSIIQFPNTYLEGSVRVVTTFDSEYCRHGTCERSEAGVCVSTSG
RWVLNNDYYRSLPGVFCGVDAVNLLTNMFTPLIQPIGALDISASIVAGGIVAIVVTCL
AYYFMRFRRAFGEYSHVVAFNTLLFLMSFTVLCLTPVYSFLPGVYSVIYLYLTFYLTN
DVSFLAHIQWMVMFTPLVPFWITIAYIICISTKHFYWFFSNYLKRRVVFNGVSFSTFE
EAALCTFLLNKEMYLKLRSDVLLPLTQYNRYLALYNKYKYFSGAMDTTSYREAACCHL
AKALNDFSNSGSDVLYQPPQTSITSAVLQSGFRKMAFPSGKVEGCMVQVTCGTTTLNG
LWLDDVVYCPRHVICTSEDMLNPNYEDLLIRKSNHNFLVQAGNVQLRVIGHSMQNCVL
KLKVDTANPKTPKYKFVRIQPGQTFSVLACYNGSPSGVYQCAMRPNFTIKGSFLNGSC
GSVGFNIDYDCVSFCYMHHMELPTGVHAGTDLEGNFYGPFVDRQTAQAAGTDTTITVN
VLAWLYAAVINGDRWFLNRFTTTLNDFNLVAMKYNYEPLTQDHVDILGPLSAQTGIAV

LDMCASLKELLQNGMNGRTILGSALLEDEFTPFDVVRQCSGVTFQSAVKRTIKGTHHW
LLLTILTSLLVLVQSTQWSLFFFLYENAFLPFAMGIIAMSAFAMMFVKHKHAFLCLFL
LPSLATVAYFNMVYMPASWVMRIMTWLDMVDTSLKLKDCVMYASAVVLLILMTARTVY
DDGARRVWTLMNVLTLVYKVYYGNALDQAISMWALIISVTSNYSGVVTTVMFLARGIV
FMCVEYCPIFFITGNTLQCIMLVYCFLGYFCTCYFGLFCLLNRYFRLTLGVYDYLVST
QEFRYMNSQGLLPPKNSIDAFKLNIKLLGVGGKPCIKVATVQSKMSDVKCTSVVLLSV
LQQLRVESSSKLWAQCVQLHNDILLAKDTTEAFEKMVSLLSVLLSMQGAVDINKLCEE
MLDNRATLQAIASEFSSLPSYAAFATAQEAYEQAVANGDSEVVLKKLKKSLNVAKSEF
DRDAAMQRKLEKMADQAMTQMYKQARSEDKRAKVTSAMQTMLFTMLRKLDNDALNNII
NNARDGCVPLNIIPLTTAAKLMVVIPDYNTYKNTCDGTTFTYASALWEIQQVVDADSK
IVQLSEISMDNSPNLAWPLIVTALRANSAVKLQNNELSPVALRQMSCAAGTTQTACTD
DNALAYYNTTKGGRFVLALLSDLQDLKWARFPKSDGTGTIYTELEPPCRFVTDTPKGP
KVKYLYFIKGLNNLNRGMVLGSLAATVRLQAGNATEVPANSTVLSFCAFAVDAAKAYK
DYLASGGQPITNCVKMLCTHTGTGQAITVTPEANMDQESFGGASCCLYCRCHIDHPNP
KGFCDLKGKYVQIPTTCANDPVGFTLKNTVCTVCGMWKGYGCSCDQLREPMLQSADAQ
SFLNRVCGVSAARLTPCGTGTSTDVVYRAFDIYNDKVAGFAKFLKTNCCRFQEKDEDD
NLIDSYFVVKRHTFSNYQHEETIYNLLKDCPAVAKHDFFKFRIDGDMVPHISRQRLTK
YTMADLVYALRHFDEGNCDTLKEILVTYNCCDDDYFNKKDWYDFVENPDILRVYANLG
ERVRQALLKTVQFCDAMRNAGIVGVLTLDNQDLNGNWYDFGDFIQTTPGSGVPVVDSY
YSLLMPILTLTRALTAESHVDTDLTKPYIKWDLLKYDFTEERLKLFDRYFKYWDQTYH
PNCVNCLDDRCILHCANFNVLFSTVFPLTSFGPLVRKIFVDGVPFVVSTGYHFRELGV
VHNQDVNLHSSRLSFKELLVYAADPAMHAASGNLLLDKRTTCFSVAALTNNVAFQTVK
PGNFNKDFYDFAVSKGFFKEGSSVELKHFFFAQDGNAAISDYDYYRYNLPTMCDIRQL
LFVVEVVDKYFDCYDGGCINANQVIVNNLDKSAGFPFNKWGKARLYYDSMSYEDQDAL
FAYTKRNVIPTITQMNLKYAISAKNRARTVAGVSICSTMTNRQFHQKLLKSIAATRGA
TVVIGTSKFYGGWHNMLKTVYSDVENPHLMGWDYPKCDRAMPNMLRIMASLVLARKHT
TCCSLSHRFYRLANECAQVLSEMVMCGGSLYVKPGGTSSGDATTAYANSVFNICQAVT
ANVNALLSTDGNKIADKYVRNLQHRLYECLYRNRDVDTDFVNEFYAYLRKHFSMMILS
DDAVVCFNSTYASQGLVASIKNFKSVLYYQNNVFMSEAKCWTETDLTKGPHEFCSQHT
MLVKQGDDYVYLPYPDPSRILGAGCFVDDIVKTDGTLMIERFVSLAIDAYPLTKHPNQ
EYADVFHLYLQYIRKLHDELTGHMLDMYSVMLTNDNTSRYWEPEFYEAMYTPHTVLQA
VGACVLCNSQTSLRCGACIRRPFLCCKCCYDHVISTSHKLVLSVNPYVCNAPGCDVTD

VTQLYLGGMSYYCKSHKPPISFPLCANGQVFGLYKNTCVGSDNVTDFNAIATCDWTNA
GDYILANTCTERLKLFAAETLKATEETFKLSYGIATVREVLSDRELHLSWEVGKPRPP
LNRNYVFTGYRVTKNSKVQIGEYTFEKGDYGDAVVYRGTTTYKLNVGDYFVLTSHTVM
PLSAPTLVPQEHYVRITGLYPTLNISDEFSSNVANYQKVGMQKYSTLQGPPGTGKSHF
AIGLALYYPSARIVYTACSHAAVDALCEKALKYLPIDKCSRIIPARARVECFDKFKVN
STLEQYVFCTVNALPETTADIVVFDEISMATNYDLSVVNARLRAKHYVYIGDPAQLPA
PRTLLTKGTLEPEYFNSVCRLMKTIGPDMFLGTCRRCPAEIVDTVSALVYDNKLKAHK
DKSAQCFKMFYKGVITHDVSSAINRPQIGVVREFLTRNPAWRKAVFISPYNSQNAVAS
KILGLPTQTVDSSQGSEYDYVIFTQTTETAHSCNVNRFNVAITRAKVGILCIMSDRDL
YDKLQFTSLEIPRRNVATLQAENVTGLFKDCSKVITGLHPTQAPTHLSVDTKFKTEGL
CVDIPGIPKDMTYRRLISMMGFKMNYQVNGYPNMFITREEAIRHVRAWIGFDVEGCHA
TREAVGTNLPLQLGFSTGVNLVAVPTGYVDTPNNTDFSRVSAKPPPGDQFKHLIPLMY
KGLPWNVVRIKIVQMLSDTLKNLSDRVVFVLWAHGFELTSMKYFVKIGPERTCCLCDR
RATCFSTASDTYACWHHSIGFDYVYNPFMIDVQQWGFTGNLQSNHDLYCQVHGNAHVA
SCDAIMTRCLAVHECFVKRVDWTIEYPIIGDELKINAACRKVQHMVVKAALLADKFPV
LHDIGNPKAIKCVPQADVEWKFYDAQPCSDKAYKIEELFYSYATHSDKFTDGVCLFWN
CNVDRYPANSIVCRFDTRVLSNLNLPGCDGGSLYVNKHAFHTPAFDKSAFVNLKQLPF
FYYSDSPCESHGKQVVSDIDYVPLKSATCITRCNLGGAVCRHHANEYRLYLDAYNMMI
SAGFSLWVYKQFDTYNLWNTFTRLQSLENVAFNVVNKGHFDGQQGEVPVSIINNTVYT
KVDGVDVELFENKTTLPVNVAFELWAKRNIKPVPEVKILNNLGVDIAANTVIWDYKRD
APAHISTIGVCSMTDIAKKPTETICAPLTVFFDGRVDGQVDLFRNARNGVLITEGSVK
GLQPSVGPKQASLNGVTLIGEAVKTQFNYYKKVDGVVQQLPETYFTQSRNLQEFKPRS
QMEIDFLELAMDEFIERYKLEGYAFEHIVYGDFSHSQLGGLHLLIGLAKRFKESPFEL
EDFIPMDSTVKNYFITDAQTGSSKCVCSVIDLLLDDFVEIIKSQDLSVVSKVVKVTID
YTEISFMLWCKDGHVETFYPKLQSSQAWQPGVAMPNLYKMQRMLLEKCDLQNYGDSAT
LPKGIMMNVAKYTQLCQYLNTLTLAVPYNMRVIHFGAGSDKGVAPGTAVLRQWLPTGT
LLVDSDLNDFVSDADSTLIGDCATVHTANKWDLIISDMYDPKTKNVTKENDSKEGFFT
YICGFIQQKLALGGSVAIKITEHSWNADLYKLMGHFAWWTAFVTNVNASSSEAFLIGC
NYLGKPREQIDGYVMHANYIFWRNTNPIQLSSYSLFDMSKFPLKLRGTAVMSLKEGQI
NDMILSLLSKGRLIIRENNRVVISSDVLVNN"

2. Download nucleotide sequence for BS001137 in fasta format and write a python code to count length and occurrence of A, C, G, T and AAC

```
input_file=open('BS001137','r')
from Bio import SeqIO
for cur_record in SeqIO.parse(input_file, "fasta"):
    #count nucleotides in this record...
    gene_name = cur_record.name
    A_count = cur_record.seq.count('A')
    C_count = cur_record.seq.count('C')
    G_count = cur_record.seq.count('G')
    T_count = cur_record.seq.count('T')
    AAC_count = cur_record.seq.count('AAC')
length = len(cur_record.seq)
#cg_percentage = float(C_count + G_count) / length
output_line = '%s\t%i\t%i\t%i\t%i\t%i\t\t%i\n' % \
(gene_name, A_count, C_count, G_count, T_count,AAC_count, length)
print("Gene name\tA_count\tC_count\tG_count\tT_count\tAAC_count\tlength")
print(output_line)
input_file.close()
```

OUTPUT

Gene name	A_count	C_count	G_count	T_count	AAC_count	length
BS001137.1	8946	5475	5853	9589	615	29863

BIOINFORMATICS
TA - 2

Q1

a. Accession number- BS001137
b. Definition: severe acute respiratory syndrome coronavirus 2 hCoV-19/Japan/KH8791/2021 RNA, complete genome.
c. source: severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2)
d. Reference: 1

Q3 1. Backbone atoms: 405, 410
Sidechain Atoms: 408, 411, 412, 413

2. N describes that atom name is Nitrogen, A describes chain identifier, "1.084", "1.00", "0.00", are x, y, z coordinator respectively.

Atharva Paliwal B40

3. Following information is extracted from 1JKZ.PDB.

[...]

ATOM 404 N ALA A 28 1.084 7.614 2.493 1.00 0.00 N

ATOM 405 CA ALA A 28 0.164 7.660 3.616 1.00 0.00 C

ATOM 406 C ALA A 28 0.842 7.090 4.856 1.00 0.00 C

ATOM 407 O ALA A 28 0.731 5.902 5.139 1.00 0.00 O

ATOM 408 CB ALA A 28 -1.123 6.911 3.287 1.00 0.00 C

ATOM 409 H ALA A 28 1.535 6.768 2.288 1.00 0.00 H

ATOM 410 HA ALA A 28 -0.085 8.696 3.802 1.00 0.00 H

ATOM 411 1HB ALA A 28 -1.278 6.918 2.218 1.00 0.00 H

ATOM 412 2HB ALA A 28 -1.957 7.396 3.773 1.00 0.00 H

ATOM 413 3HB ALA A 28 -1.047 5.891 3.634 1.00 0.00 H

[...]

1. Which atom numbers correspond to the backbone atoms and which atomnumbers correspond to the sidechain of this amino acid ?

→Backbone Atoms: 405,410

Side chain atoms: 408,411,412,413

2. Describe the information in the following columns (indicated by the values in the first record): " N ", "A", "1.084", "1.00", "0.00".

→N describes that atom name is Nitrogen, A describes chain identifier, "1.084", "1.00", "0.00" are x, y, z coordinates respectively.

4. Write biopython code to read the file sequence_insulin.fasta. Acquire the sequences and perform the following operations on the sequences present.

1) Complement the sequence name it Seq1

2) Reverse complement the Seq1

3) Generate mRNA sequence from DNA sequence.

4) Find the Protein sequence from the DNA sequence. Use all the available codon tables and print the output.

5) Find out the GC rich sequence from the available sequences.

```
print("hello")

# Converts sequence to rNA

def convert_sequence(sequence):  # Takes sequence and type of secuence
    # if the sequence is DNA: convert t to u
    #conversion_dict = {
     #   'A': 'U',
      #  'T': 'A',
       # 'C': 'G',
        #'G': 'C'
    #}
    conversion_dict = {
        'A': 'A',
        'T': 'U',
        'C': 'C',
        'G': 'G'
    }
 # convert sequence into a list
    converted_sequence = []
    sequence_list = list(sequence)

    # convert list one by one, checking the dictionary for the corresponding key, and add it to the new clist

 for i in sequence_list:
        converted_sequence.append(conversion_dict[i])

    # return converted sequence, seperated by a space every three spaces
    converted_sequence = ''.join(converted_sequence)
    # noinspection PyTypeChecker
    return ' '.join([converted_sequence[i:i + 3] for i in range(0, len(converted_sequence), 3)])


def convert_dna_to_protein(dna_seq):
    codon_table = {
```

```
    'ATA':'I', 'ATC':'I', 'ATT':'I', 'ATG':'M',
    'ACA':'T', 'ACC':'T', 'ACG':'T', 'ACT':'T',
    'AAC':'N', 'AAT':'N', 'AAA':'K', 'AAG':'K',
    'AGC':'S', 'AGT':'S', 'AGA':'R', 'AGG':'R',
    'CTA':'L', 'CTC':'L', 'CTG':'L', 'CTT':'L',
    'CCA':'P', 'CCC':'P', 'CCG':'P', 'CCT':'P',
    'CAC':'H', 'CAT':'H', 'CAA':'Q', 'CAG':'Q',
    'CGA':'R', 'CGC':'R', 'CGG':'R', 'CGT':'R',
    'GTA':'V', 'GTC':'V', 'GTG':'V', 'GTT':'V',
    'GCA':'A', 'GCC':'A', 'GCG':'A', 'GCT':'A',
    'GAC':'D', 'GAT':'D', 'GAA':'E', 'GAG':'E',
    'GGA':'G', 'GGC':'G', 'GGG':'G', 'GGT':'G',
    'TCA':'S', 'TCC':'S', 'TCG':'S', 'TCT':'S',
    'TTC':'F', 'TTT':'F', 'TTA':'L', 'TTG':'L',
    'TAC':'Y', 'TAT':'Y', 'TAA':'_', 'TAG':'_',
    'TGC':'C', 'TGT':'C', 'TGA':'_', 'TGG':'W',
  }
  protein_seq =""
  print(len(dna_seq))
  while(len(dna_seq)%3!=0):
    dna_seq=dna_seq[:-1]

  if len(dna_seq)%3 == 0:
    for i in range(0, len(dna_seq), 3):
      codon = dna_seq[i:i + 3]
      protein_seq+= codon_table[codon]
      #print(protein_seq)
  return protein_seq

input_file=open('D:\\Monica\\7th sem\\Bio
Informatics\\TA2\\sequence_insulin.fasta','r')
from Bio import SeqIO
for cur_record in SeqIO.parse(input_file, "fasta"):
  print(cur_record)
  #print(cur_record.seq.complement())
  seq1=cur_record.seq.complement()
  rev_seq1=seq1.reverse_complement()
  print("Complemented Seq1: ",seq1)
  print("\nReverse Complemented seq1: ",rev_seq1)
  print("\n\n")

  mRNA=convert_sequence(cur_record.seq)
  print(mRNA)

protein_seq=convert_dna_to_protein(cur_record.seq)
  print("\nProtein String: ",protein_seq)

  C_count = cur_record.seq.count('C')
  G_count = cur_record.seq.count('G')
  length = len(cur_record.seq)
  cg_percentage = ((C_count + G_count) /
length )*100
  print("GC Rich % is ",cg_percentage)
print("\n\n")
```

OUTPUT

ID: NM_000207.3

Name: NM_000207.3

Description: NM_000207.3 Homo sapiens insulin (INS), transcript variant 1, mRNA

Number of features: 0

Seq('AGCCCTCCAGGACAGGCTGCATCAGAAG
AGGCCATCAAGCAGATCACTGTCCTT...AGC')

Complemented Seq1:
TCGGGAGGTCCTGTCCGACGTAGTCTTCTCCG

GTAGTTCGTCTAGTGACAGGAAGACGGTACC
GGGACACCTACGCGGAGGACGGGGACGACC
GCGACGACCGGGAGACCCCTGGACTGGGTCG
GCGTCGGAAACACTTGGTTGTGGACACGCCG
AGTGTGGACCACCTTCGAGAGATGGATCACA
CGCCCCTTGCTCCGAAGAAGATGTGTGGGTT
CTGGGCGGCCCTCCGTCTCCTGGACGTCCAC
CCCGTCCACCTCGACCCGCCCCCGGGACCAC
GTCCGTCGGACGTCGGGAACCGGGACCTCCC
CAGGGACGTCTTCGCACCGTAACACCTTGTT
ACGACATGGTCGTAGACGAGGGAGATGGTCG
ACCTCTTGATGACGTTGATCTGCGTCGGGCGT
CCGTCGGGGTGTGGGCGGCGGAGGACGTGGC
TCTCTCTACCTTATTTCGGGAACTTGGTCG

Reverse Complemented seq1:
CGACCAAGTTCCCGAAATAAGGTAGAGAGA
GCCACGTCCTCCGCCGCCCACACCCCGACGG
ACGCCCGACGCAGATCAACGTCATCAAGAGG
TCGACCATCTCCCTCGTCTACGACCATGTCGT
AACAAGGTGTTACGGTGCGAAGACGTCCCTG
GGGAGGTCCCGGTTCCCGACGTCCGACGGAC
GTGGTCCCGGGGGCGGGTCGAGGTGGACGG
GGTGGACGTCCAGGAGACGGAGGGCCGCCC
AGAACCCACACATCTTCTTCGGAGCAAGGGG
CGTGTGATCCATCTCTCGAAGGTGGTCCACA
CTCGGCGTGTCCACAACCAAGTGTTTCCGAC
GCCGACCCAGTCCAGGGGTCTCCCGGTCGTC
GCGGTCGTCCCCGTCCTCCGCGTAGGTGTCCC
GGTACCGTCTTCCTGTCACTAGACGAACTAC
CGGAGAAGACTACGTCGGACAGGACCTCCCG
A

AGC CCU CCA GGA CAG GCU GCA UCA GAA
GAG GCC AUC AAG CAG AUC ACU GUC CUU
CUG CCA UGG CCC UGU GGA UGC GCC UCC
UGC CCC UGC UGG CGC UGC UGG CCC UCU
GGG GAC CUG ACC CAG CCG CAG CCU UUG
UGA ACC AAC ACC UGU GCG GCU CAC ACC
UGG UGG AAG CUC UCU ACC UAG UGU GCG
GGG AAC GAG GCU UCU UCU ACA CAC CCA
AGA CCC GCC GGG AGG CAG AGG ACC UGC
AGG UGG GGC AGG UGG AGC UGG GCG GGG
GCC CUG GUG CAG GCA GCC UGC AGC CCU
UGG CCC UGG AGG GGU CCC UGC AGA AGC
GUG GCA UUG UGG AAC AAU GCU GUA CCA
GCA UCU GCU CCC UCU ACC AGC UGG AGA
ACU ACU GCA ACU AGA CGC AGC CCG CAG
GCA GCC CCA CAC CCG CCG CCU CCU GCA
CCG AGA GAG AUG GAA UAA AGC CCU UGA
ACC AGC

465

Protein String:
SPPGQAASEEAIKQITVLLPWPCGCASCPCWRC
WPSGDLTQPQPL_TNTCAAHTWWKLST_CAGN
EASSTHPRPAGRQRTCRWGRWSWAGALVQAA
CSPWPWRGPCRSVALWNNAVPASAPSTSWRT
TATRRSPQAAPHPPPPAPREME_SP_TS

GC Rich % is 63.87096774193548

ID: NM_001185097.2

Name: NM_001185097.2

Description: NM_001185097.2 Homo sapiens insulin (INS), transcript variant 2, mRNA

Number of features: 0

Seq('AGCCCTCCAGGACAGGCTGCATCAGAAG
AGGCCATCAAGCAGGTCTGTTCCAAG...AGC')

Complemented Seq1:
TCGGGAGGTCCTGTCCGACGTAGTCTTCTCCG
GTAGTTCGTCCAGACAAGGTTCCCGGAAACG
CAGTCTAGTGACAGGAAGACGGTACCGGGAC
ACCTACGCGGAGGACGGGGACGACCGCGAC
GACCGGGAGACCCCTGGACTGGGTCGGCGTC
GGAAACACTTGGTTGTGGACACGCCGAGTGT
GGACCACCTTCGAGAGATGGATCACACGCCC
CTTGCTCCGAAGAAGATGTGTGGGTTCTGGG
CGGCCCTCCGTCTCCTGGACGTCCACCCCGTC
CACCTCGACCCGCCCCCGGGACCACGTCCGT
CGGACGTCGGGAACCGGGACCTCCCCAGGGA
CGTCTTCGCACCGTAACACCTTGTTACGACAT
GGTCGTAGACGAGGGAGATGGTCGACCTCTT
GATGACGTTGATCTGCGTCGGGCGTCCGTCG
GGGTGTGGGCGGCGGAGGACGTGGCTCTCTC
TACCTTATTTCGGGAACTTGGTCG

Reverse Complemented seq1:
CGACCAAGTTCCCGAAATAAGGTAGAGAGA
GCCACGTCCTCCGCCGCCCACACCCCGACGG

ACGCCCGACGCAGATCAACGTCATCAAGAGG
TCGACCATCTCCCTCGTCTACGACCATGTCGT
AACAAGGTGTTACGGTGCGAAGACGTCCCTG
GGGAGGTCCCGGTTCCCGACGTCCGACGGAC
GTGGTCCCGGGGGCGGGTCGAGGTGGACGG
GGTGGACGTCCAGGAGACGGAGGGCCGCCC
AGAACCCACACATCTTCTTCGGAGCAAGGGG
CGTGTGATCCATCTCTCGAAGGTGGTCCACA
CTCGGCGTGTCCACAACCAAGTGTTTCCGAC
GCCGACCCAGTCCAGGGGTCTCCCGGTCGTC
GCGGTCGTCCCCGTCCTCCGCGTAGGTGTCCC
GGTACCGTCTTCCTGTCACTAGACTGCGTTTC
CGGGAACCTTGTCTGGACGAACTACCGGAGA
AGACTACGTCGGACAGGACCTCCCGA

AGC CCU CCA GGA CAG GCU GCA UCA GAA
GAG GCC AUC AAG CAG GUC UGU UCC AAG
GGC CUU UGC GUC AGA UCA CUG UCC UUC
UGC CAU GGC CCU GUG GAU GCG CCU CCU
GCC CCU GCU GGC GCU GCU GGC CCU CUG
GGG ACC UGA CCC AGC CGC AGC CUU UGU
GAA CCA ACA CCU GUG CGG CUC ACA CCU
GGU GGA AGC UCU CUA CCU AGU GUG CGG
GGA ACG AGG CUU CUU CUA CAC ACC CAA
GAC CCG CCG GGA GGC AGA GGA CCU GCA
GGU GGG GCA GGU GGA GCU GGG CGG GGG
CCC UGG UGC AGG CAG CCU GCA GCC CUU
GGC CCU GGA GGG GUC CCU GCA GAA GCG
UGG CAU UGU GGA ACA AUG CUG UAC CAG
CAU CUG CUC CCU CUA CCA GCU GGA GAA
CUA CUG CAA CUA GAC GCA GCC CGC AGG
CAG CCC CAC ACC CGC CGC CUC CUG CAC
CGA GAG AGA UGG AAU AAA GCC CUU GAA
CCA GC

491

Protein String:
SPPGQAASEEAIKQVCSKGLCVRSLSFCHGPVD
APPAPAGAAGPLGT_PSRSLCEPTPVRLTPGGSS
LPSVRGTRLLLHTQDPPGGRGPAGGAGGAGRG
PWCRQPAALGPGGVPAEAWHCGTMLYQHLLP
LPAGELLQLDAARRQPHTRRLLHRERWNKALE
P

GC Rich % is 63.543788187372705

ID: NM_001185098.2

Name: NM_001185098.2

Description: NM_001185098.2 Homo sapiens insulin (INS), transcript variant 3, mRNA

Number of features: 0

Seq('AGCCCTCCAGGACAGGCTGCATCAGAAG
AGGCCATCAAGCAGGTCTGTTCCAAG...AGC')

Complemented Seq1:
TCGGGAGGTCCTGTCCGACGTAGTCTTCTCCG
GTAGTTCGTCCAGACAAGGTTCCCGGAAACG
CAGTCCACCCGAGTCCTAAGGTCCCACCGAC
CTGGGGTCCGGGGTCGAGACGTCGTCCCTCC
TGCACCGACCCGAGCACTTCGTACACCCCCA
CTCGGGTCCCCGGGGTTCCGTCCCGTGGACC
GGAAGTCGGACGGAGTCGGGACGGACAGAG
GGTCTAGTGACAGGAAGACGGTACCGGGAC
ACCTACGCGGAGGACGGGGACGACCGCGAC
GACCGGGAGACCCCTGGACTGGGTCGGCGTC
GGAAACACTTGGTTGTGGACACGCCGAGTGT
GGACCACCTTCGAGAGATGGATCACACGCCC
CTTGCTCCGAAGAAGATGTGTGGGTTCTGGG
CGGCCCTCCGTCTCCTGGACGTCCACCCCGTC
CACCTCGACCCGCCCCCGGGACCACGTCCGT
CGGACGTCGGGAACCGGGACCTCCCCAGGGA
CGTCTTCGCACCGTAACACCTTGTTACGACAT
GGTCGTAGACGAGGGAGATGGTCGACCTCTT
GATGACGTTGATCTGCGTCGGGCGTCCGTCG
GGGTGTGGGCGGCGGAGGACGTGGCTCTCTC
TACCTTATTTCGGGAACTTGGTCG

Reverse Complemented seq1:
CGACCAAGTTCCCGAAATAAGGTAGAGAGA
GCCACGTCCTCCGCCGCCCACACCCCGACGG
ACGCCCGACGCAGATCAACGTCATCAAGAGG
TCGACCATCTCCCTCGTCTACGACCATGTCGT
AACAAGGTGTTACGGTGCGAAGACGTCCCTG
GGGAGGTCCCGGTTCCCGACGTCCGACGGAC
GTGGTCCCGGGGGCGGGTCGAGGTGGACGG
GGTGGACGTCCAGGAGACGGAGGGCCGCCC
AGAACCCACACATCTTCTTCGGAGCAAGGGG
CGTGTGATCCATCTCTCGAAGGTGGTCCACA
CTCGGCGTGTCCACAACCAAGTGTTTCCGAC
GCCGACCCAGTCCAGGGGTCTCCCGGTCGTC
GCGGTCGTCCCCGTCCTCCGCGTAGGTGTCCC
GGTACCGTCTTCCTGTCACTAGACCCTCTGTC
CGTCCCGACTCCGTCCGACTTCCGGTCCACG
GGACGGAACCCCGGGGACCCGAGTGGGGGT
GTACGAAGTGCTCGGGTCGGTGCAGGAGGGA
CGACGTCTCGACCCCGGACCCCAGGTCGGTG
GGACCTTAGGACTCGGGTGGACTGCGTTTCC
GGGAACCTTGTCTGGACGAACTACCGGAGAA
GACTACGTCGGACAGGACCTCCCGA

AGC CCU CCA GGA CAG GCU GCA UCA GAA
GAG GCC AUC AAG CAG GUC UGU UCC AAG
GGC CUU UGC GUC AGG UGG GCU CAG GAU
UCC AGG GUG GCU GGA CCC CAG GCC CCA
GCU CUG CAG CAG GGA GGA CGU GGC UGG
GCU CGU GAA GCA UGU GGG GGU GAG CCC
AGG GGC CCC AAG GCA GGG CAC CUG GCC
UUC AGC CUG CCU CAG CCC UGC CUG UCU
CCC AGA UCA CUG UCC UUC UGC CAU GGC
CCU GUG GAU GCG CCU CCU GCC CCU GCU
GGC GCU GCU GGC CCU CUG GGG ACC UGA
CCC AGC CGC AGC CUU UGU GAA CCA ACA
CCU GUG CGG CUC ACA CCU GGU GGA AGC
UCU CUA CCU AGU GUG CGG GGA ACG AGG
CUU CUU CUA CAC ACC CAA GAC CCG CCG
GGA GGC AGA GGA CCU GCA GGU GGG GCA
GGU GGA GCU GGG CGG GGG CCC UGG UGC
AGG CAG CCU GCA GCC CUU GGC CCU GGA
GGG GUC CCU GCA GAA GCG UGG CAU UGU
GGA ACA AUG CUG UAC CAG CAU CUG CUC
CCU CUA CCA GCU GGA GAA CUA CUG CAA
CUA GAC GCA GCC CGC AGG CAG CCC CAC
ACC CGC CGC CUC CUG CAC CGA GAG AGA
UGG AAU AAA GCC CUU GAA CCA GC

644

Protein String:
SPPGQAASEEAIKQVCSKGLCVRWAQDSRVAG
PQAPALQQGGRGWAREACGGEPRGPKAGHLA
FSLPQPCLSPRSLSFCHGPVDAPPAPAGAAGPL
GT_PSRSLCEPTPVRLTPGGSSLPSVRGTRLLLH
TQDPPGGRGPAGGAGGAGRGPWCRQPAALGP
GGVPAEAWHCGTMLYQHLLPLPAGELLQLDA
ARRQPHTRRLLHRERWNKALEP

GC Rich % is 65.06211180124224

ID: NM_001291897.2

Name: NM_001291897.2

Description: NM_001291897.2 Homo sapiens insulin (INS), transcript variant 4, mRNA

Number of features: 0

Seq('AGCCCTCCAGGACAGGCTGCATCAGAAG
AGGCCATCAAGCAGGTCTGTTCCAAG...AGC')

Complemented Seq1:
TCGGGAGGTCCTGTCCGACGTAGTCTTCTCCG
GTAGTTCGTCCAGACAAGGTTCCCGGAAACG
CAGTCCACCCGAGTCCTAAGGTCCCACCGAC
CTGGGGTCTAGTGACAGGAAGACGGTACCGG
GACACCTACGCGGAGGACGGGGACGACCGC
GACGACCGGGAGACCCCTGGACTGGGTCGGC
GTCGGAAACACTTGGTTGTGGACACGCCGAG
TGTGGACCACCTTCGAGAGATGGATCACACG
CCCCTTGCTCCGAAGAAGATGTGTGGGTTCT
GGGCGGCCCTCCGTCTCCTGGACGTCCACCC
CGTCCACCTCGACCCGCCCCCGGGACCACGT
CCGTCGGACGTCGGGAACCGGGACCTCCCCA
GGGACGTCTTCGCACCGTAACACCTTGTTAC
GACATGGTCGTAGACGAGGGAGATGGTCGAC
CTCTTGATGACGTTGATCTGCGTCGGGCGTCC
GTCGGGGTGTGGGCGGCGGAGGACGTGGCTC
TCTCTACCTTATTTCGGGAACTTGGTCG

Reverse Complemented seq1:
CGACCAAGTTCCCGAAATAAGGTAGAGAGA
GCCACGTCCTCCGCCGCCCACACCCCGACGG
ACGCCCGACGCAGATCAACGTCATCAAGAGG
TCGACCATCTCCCTCGTCTACGACCATGTCGT
AACAAGGTGTTACGGTGCGAAGACGTCCCTG
GGGAGGTCCCGGTTCCCGACGTCCGACGGAC
GTGGTCCCGGGGGCGGGTCGAGGTGGACGG
GGTGGACGTCCAGGAGACGGAGGGCCGCCC
AGAACCCACACATCTTCTTCGGAGCAAGGGG
CGTGTGATCCATCTCTCGAAGGTGGTCCACA
CTCGGCGTGTCCACAACCAAGTGTTTCCGAC
GCCGACCCAGTCCAGGGGTCTCCCGGTCGTC
GCGGTCGTCCCCGTCCTCCGCGTAGGTGTCCC
GGTACCGTCTTCCTGTCACTAGACCCCAGGTC
GGTGGGACCTTAGGACTCGGGTGGACTGCGT
TTCCGGGAACCTTGTCTGGACGAACTACCGG
AGAAGACTACGTCGGACAGGACCTCCCGA

AGC CCU CCA GGA CAG GCU GCA UCA GAA
GAG GCC AUC AAG CAG GUC UGU UCC AAG
GGC CUU UGC GUC AGG UGG GCU CAG GAU
UCC AGG GUG GCU GGA CCC CAG AUC ACU
GUC CUU CUG CCA UGG CCC UGU GGA UGC
GCC UCC UGC CCC UGC UGG CGC UGC UGG
CCC UCU GGG GAC CUG ACC CAG CCG CAG
CCU UUG UGA ACC AAC ACC UGU GCG GCU
CAC ACC UGG UGG AAG CUC UCU ACC UAG
UGU GCG GGG AAC GAG GCU UCU UCU ACA
CAC CCA AGA CCC GCC GGG AGG CAG AGG
ACC UGC AGG UGG GGC AGG UGG AGC UGG
GCG GGG GCC CUG GUG CAG GCA GCC UGC
AGC CCU UGG CCC UGG AGG GGU CCC UGC

AGA AGC GUG GCA UUG UGG AAC AAU GCU GUA CCA GCA UCU GCU CCC UCU ACC AGC UGG AGA ACU ACU GCA ACU AGA CGC AGC CCG CAG GCA GCC CCA CAC CCG CCG CCU CCU GCA CCG AGA GAG AUG GAA UAA AGC CCU UGA ACC AGC

525

Protein String:
SPPGQAASEEAIKQVCSKGLCVRWAQDSRVAG PQITVLLPWPCGCASCPCWRCWPSGDLTQPQP L_TNTCAAHTWWKLST_CAGNEASSTHPRPAG RQRTCRWGRWSWAGALVQAACSPWPWRGPC RSVALWNNAVPASAPSTSWRTTATRRSPQAAP HPPPPAPREME_SP_TS

GC Rich % is 63.8095238095238

5. Analyze the occurrence of similar proteins in "nr" and SWISS-PROT database

for the sequence given below:

>1336093|Genbank|Outer membrane integral membrane protein|HrcC

MVEKRELRCRLLGALLMLCATLPAGAQTPADWKEQSYAYSADRTPLSTVLQDFADGH

SVD

LHLGNVEDTEVTAKIRAENASAFLDRLALEHHFQWFVYNNTLYVSPQDEQSSERLEISP

D

AAPDIKQALSGIGLLDPRFGWGELPDDGVVLVTGPPQYLELVKRFSEQREKKEDRRKV

MT

FPLRYASVADRTIHYRDQTVVIPGVATMLNELMNGKRAAPASASGIDSTPGGPDTNSM

MQ

NTQTLLSRLSSRNKTSNRAGGRDNEIEDVSGRISADVRNNALLIRDDDKRHDEYSQLIAK

IDVPQNLVEIDAVILDIDRTALNRLEANWQATLGGVTGGSSLMSGSGTLFVSDFKRFFAD

IQALEGEGTASIVANPSVLTLENQPAVIDFSQTAYITATGERVADIQPVTAGTSLQVTPR

AVGNEGHSSIQLMIDIEDGHVQTNGDGQATGVKRGTVSTQALISENRALVLGGFHVEES

A

DRDRRIPLLGDIPWLGQLFSSKRHEISQRQRLFILTPRLIGDQTDPTRYVTADNRQQLSDA

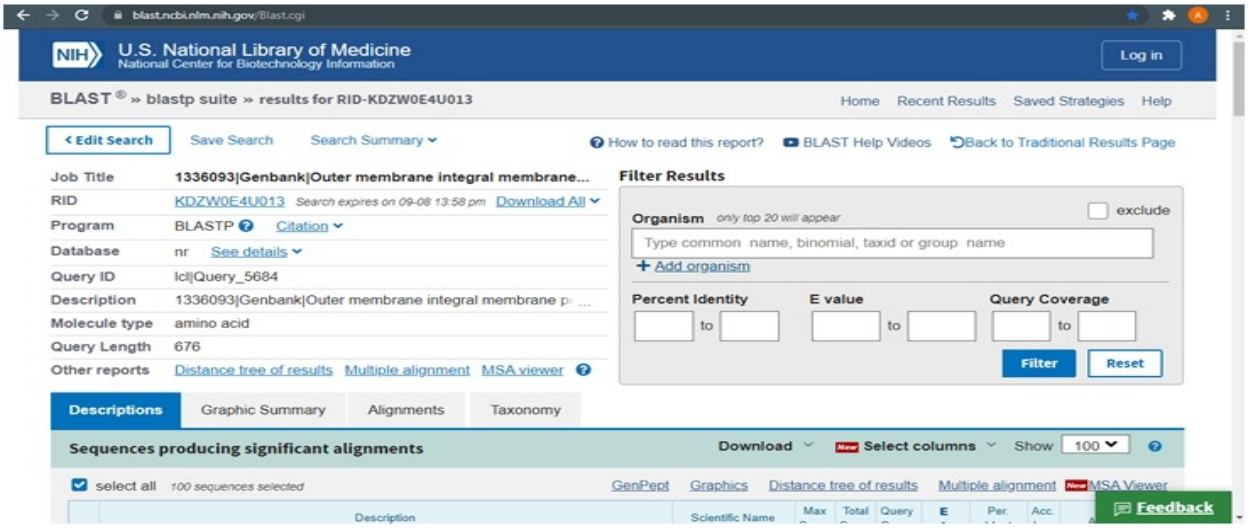
MGRVERRHSSVNQHDVVENALRDLAEGQSPAGFQPQTSGTRLSEVCRSTPALLFESTRG

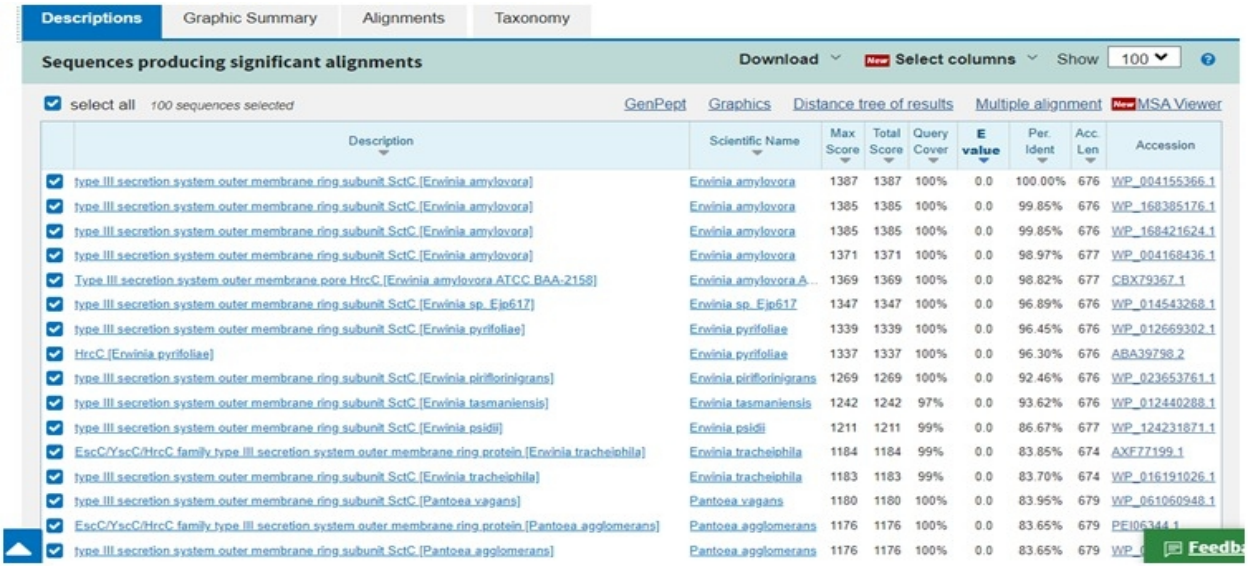
QWYSSSTNGVQLSVGVVRNTSSKPLRFDEANCASKRTLAVAVWPHSALAPGESAEVYL

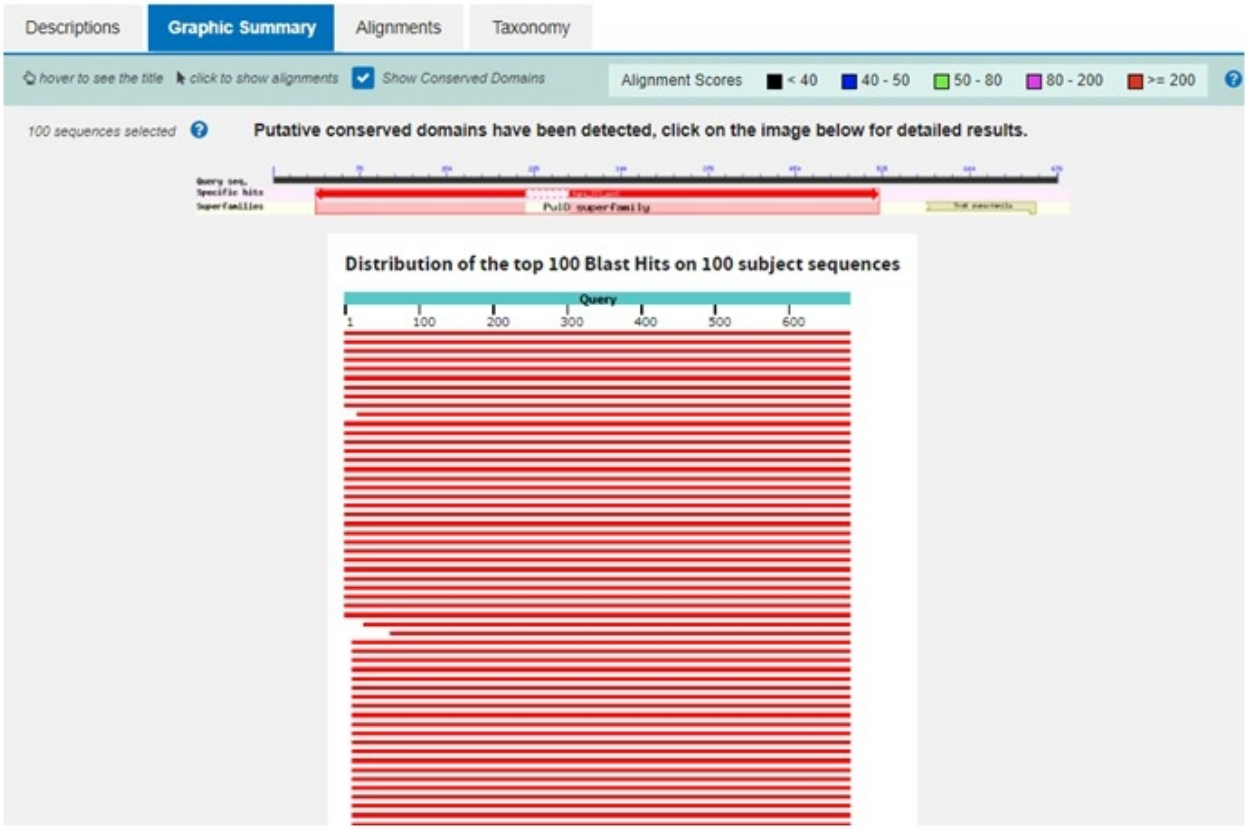
AM

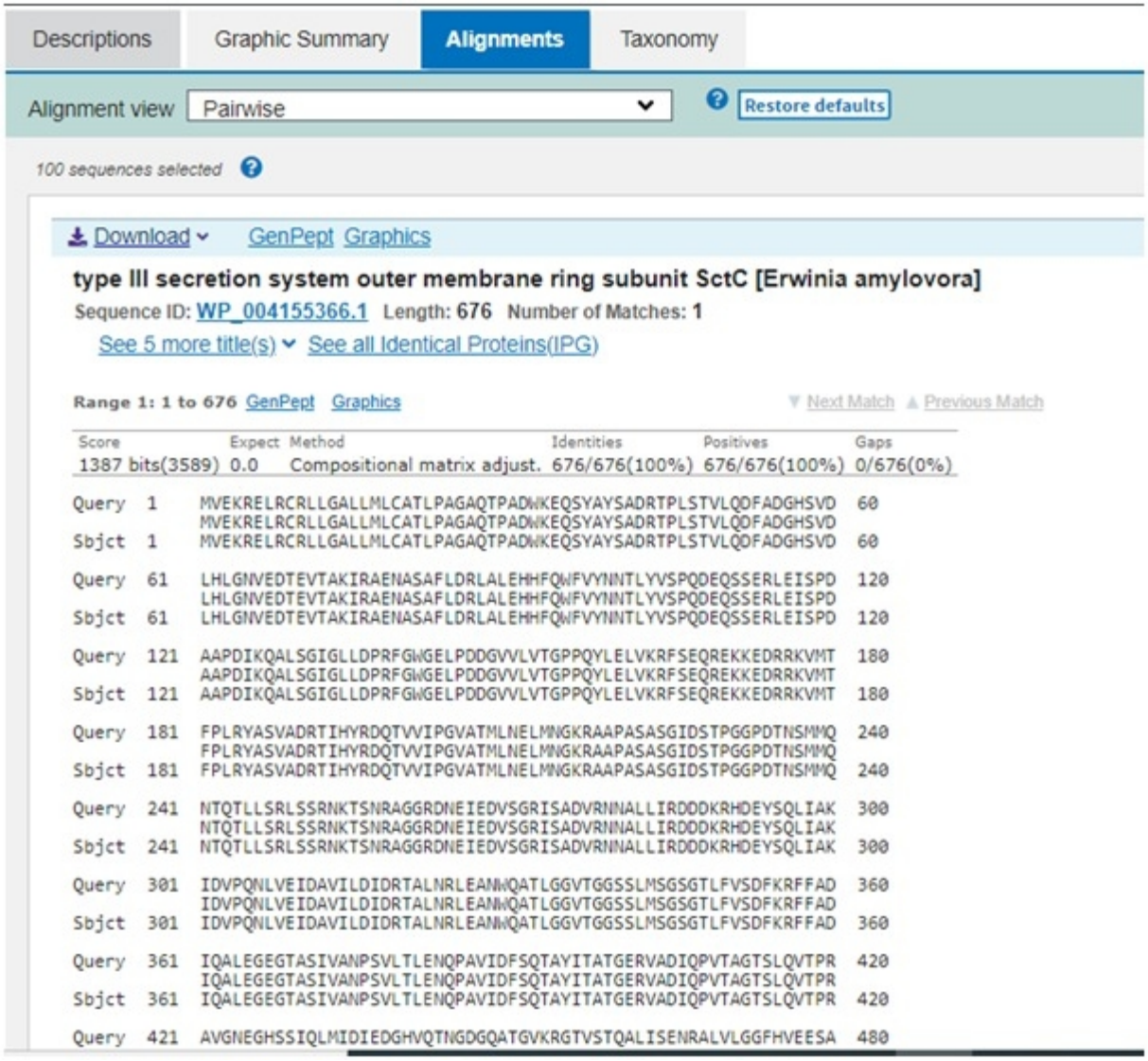
DPSRVLHASRESLLNR

(NOTE: Use blast tool)


blast.ncbi.nlm.nih.gov/Blast.cgi
U.S. National Library of Medicine
National Center for Biotechnology Information
Log in
BLAST® » blastp suite » results for RID-KDZW0E4U013
Home Recent Results Saved Strategies Help
< Edit Search
Save Search
Search Summary
How to read this report?
BLAST Help Videos
Back to Traditional Results Page
Job Title 1336093|Genbank|Outer membrane integral membrane...
RID KDZW0E4U013 Search expires on 09-08 13:58 pm Download All
Program BLASTP Citation
Database nr See details
Query ID lcl|Query_5684
Description 1336093|Genbank|Outer membrane integral membrane p
Molecule type amino acid
Query Length 676
Other reports Distance tree of results Multiple alignment MSA viewer
Filter Results
Organism only top 20 will appear
exclude
Type common name, binomial, taxid or group name
Add organism
Percent Identity
E value
Query Coverage
to
Filter
Reset
Descriptions
Graphic Summary
Alignments
Taxonomy
Sequences producing significant alignments
Download
Select columns
Show 100
select all 100 sequences selected
GenPept
Graphics
Distance tree of results
Multiple alignment
MSA Viewer
Description
Scientific Name
Feedback


Descriptions
Graphic Summary
Alignments
Taxonomy
Sequences producing significant alignments
Download
Select columns
Show 100
select all 100 sequences selected
GenPept
Graphics
Distance tree of results
Multiple alignment
MSA Viewer
Description
Scientific Name
Max Score
Total Score
Query Cover
E value
Per. Ident
Acc. Len
Accession
type III secretion system outer membrane ring subunit SctC [Erwinia amylovora] Erwinia amylovora 1387 1387 100% 0.0 100.00% 676 WP_004155366.1
type III secretion system outer membrane ring subunit SctC [Erwinia amylovora] Erwinia amylovora 1385 1385 100% 0.0 99.85% 676 WP_168385176.1
type III secretion system outer membrane ring subunit SctC [Erwinia amylovora] Erwinia amylovora 1385 1385 100% 0.0 99.85% 676 WP_168421624.1
type III secretion system outer membrane ring subunit SctC [Erwinia amylovora] Erwinia amylovora 1371 1371 100% 0.0 98.97% 677 WP_004168436.1
Type III secretion system outer membrane pore HrcC [Erwinia amylovora ATCC BAA-2158] Erwinia amylovora A... 1369 1369 100% 0.0 98.82% 677 CBX79367.1
type III secretion system outer membrane ring subunit SctC [Erwinia sp. Ejp617] Erwinia sp. Ejp617 1347 1347 100% 0.0 96.89% 676 WP_014543268.1
type III secretion system outer membrane ring subunit SctC [Erwinia pyrifoliae] Erwinia pyrifoliae 1339 1339 100% 0.0 96.45% 676 WP_012669302.1
HrcC [Erwinia pyrifoliae] Erwinia pyrifoliae 1337 1337 100% 0.0 96.30% 676 ABA39798.2
type III secretion system outer membrane ring subunit SctC [Erwinia piriflorinigrans] Erwinia piriflorinigrans 1269 1269 100% 0.0 92.46% 676 WP_023653761.1
type III secretion system outer membrane ring subunit SctC [Erwinia tasmaniensis] Erwinia tasmaniensis 1242 1242 97% 0.0 93.62% 676 WP_012440288.1
type III secretion system outer membrane ring subunit SctC [Erwinia psidii] Erwinia psidii 1211 1211 99% 0.0 86.67% 677 WP_124231871.1
EscC/YscC/HrcC family type III secretion system outer membrane ring protein [Erwinia tracheiphila] Erwinia tracheiphila 1184 1184 99% 0.0 83.85% 674 AXF77199.1
type III secretion system outer membrane ring subunit SctC [Erwinia tracheiphila] Erwinia tracheiphila 1183 1183 99% 0.0 83.70% 674 WP_016191026.1
type III secretion system outer membrane ring subunit SctC [Pantoea vagans] Pantoea vagans 1180 1180 100% 0.0 83.95% 679 WP_061060948.1
EscC/YscC/HrcC family type III secretion system outer membrane ring protein [Pantoea agglomerans] Pantoea agglomerans 1176 1176 100% 0.0 83.65% 679 PEI06344.1
type III secretion system outer membrane ring subunit SctC [Pantoea agglomerans] Pantoea agglomerans 1176 1176 100% 0.0 83.65% 679 WP_
Feedb


Descriptions
Graphic Summary
Alignments
Taxonomy
hover to see the title
click to show alignments
Show Conserved Domains
Alignment Scores
< 40
40 - 50
50 - 80
80 - 200
>= 200
100 sequences selected
Putative conserved domains have been detected, click on the image below for detailed results.
Query seq.
Specific hits
Superfamilies
PulD superfamily
Distribution of the top 100 Blast Hits on 100 subject sequences
Query
1
100
200
300
400
500
600


Descriptions
Graphic Summary
Alignments
Taxonomy
Alignment view
Pairwise
Restore defaults
100 sequences selected
Download
GenPept
Graphics
type III secretion system outer membrane ring subunit SctC [Erwinia amylovora]
Sequence ID: WP_004155366.1 Length: 676 Number of Matches: 1
See 5 more title(s)
See all Identical Proteins(IPG)
Range 1: 1 to 676 GenPept Graphics
Next Match
Previous Match
Score Expect Method Identities Positives Gaps
1387 bits(3589) 0.0 Compositional matrix adjust. 676/676(100%) 676/676(100%) 0/676(0%)
Query 1 MVEKRELRCRLLGALLMLCATLPAGAQTPADWKEQSYAYSADRTPLSTVLQDFADGHSVD 60
MVEKRELRCRLLGALLMLCATLPAGAQTPADWKEQSYAYSADRTPLSTVLQDFADGHSVD
Sbjct 1 MVEKRELRCRLLGALLMLCATLPAGAQTPADWKEQSYAYSADRTPLSTVLQDFADGHSVD 60
Query 61 LHLGNVEDTEVTAKIRAENASAFLDRLALEHHFQWFVYNNTLYVSPQDEQSSERLEISPD 120
LHLGNVEDTEVTAKIRAENASAFLDRLALEHHFQWFVYNNTLYVSPQDEQSSERLEISPD
Sbjct 61 LHLGNVEDTEVTAKIRAENASAFLDRLALEHHFQWFVYNNTLYVSPQDEQSSERLEISPD 120
Query 121 AAPDIKQALSGIGLLDPRFGWGELPDDGVVLVTGPPQYLELVKRFSEQREKKEDRRKVMT 180
AAPDIKQALSGIGLLDPRFGWGELPDDGVVLVTGPPQYLELVKRFSEQREKKEDRRKVMT
Sbjct 121 AAPDIKQALSGIGLLDPRFGWGELPDDGVVLVTGPPQYLELVKRFSEQREKKEDRRKVMT 180
Query 181 FPLRYASVADRTIHYRDQTVVIPGVATMLNELMNGKRAAPASASGIDSTPGGPDTNSMMQ 240
FPLRYASVADRTIHYRDQTVVIPGVATMLNELMNGKRAAPASASGIDSTPGGPDTNSMMQ
Sbjct 181 FPLRYASVADRTIHYRDQTVVIPGVATMLNELMNGKRAAPASASGIDSTPGGPDTNSMMQ 240
Query 241 NTQTLLSRLSSRNKTSNRAGGRDNEIEDVSGRISADVRNNALLIRDDDKRHDEYSQLIAK 300
NTQTLLSRLSSRNKTSNRAGGRDNEIEDVSGRISADVRNNALLIRDDDKRHDEYSQLIAK
Sbjct 241 NTQTLLSRLSSRNKTSNRAGGRDNEIEDVSGRISADVRNNALLIRDDDKRHDEYSQLIAK 300
Query 301 IDVPQNLVEIDAVILDIDRTALNRLEAMMQATLGGVTGGSSLMSGSGTLFVSDFKRFFAD 360
IDVPQNLVEIDAVILDIDRTALNRLEAMMQATLGGVTGGSSLMSGSGTLFVSDFKRFFAD
Sbjct 301 IDVPQNLVEIDAVILDIDRTALNRLEAMMQATLGGVTGGSSLMSGSGTLFVSDFKRFFAD 360
Query 361 IQALEGEGTASIVANPSVLTLENQPAVIDFSQTAYITATGERVADIQPVTAGTSLQVTPR 420
IQALEGEGTASIVANPSVLTLENQPAVIDFSQTAYITATGERVADIQPVTAGTSLQVTPR
Sbjct 361 IQALEGEGTASIVANPSVLTLENQPAVIDFSQTAYITATGERVADIQPVTAGTSLQVTPR 420
Query 421 AVGNEGHSSIQLMIDIEDGHVQTNGDGQATGVKRGTVSTQALISENRALVLGGFHVEESA 480

Descriptions | Graphic Summary | Alignments | **Taxonomy**

Reports | **Lineage** | Organism | Taxonomy

100 sequences selected

Organism	Blast Name	Score	Number of Hits	Description
Enterobacterales	enterobacteria		207	
. Erwiniaceae	enterobacteria		65	
. . Erwinia	enterobacteria		33	
. . . Erwinia amylovora	enterobacteria	1387	8	Erwinia amylovora hits
. . . Erwinia amylovora ACW56400	enterobacteria	1387	1	Erwinia amylovora ACW56400 hits
. . . Erwinia amylovora Ea644	enterobacteria	1371	1	Erwinia amylovora Ea644 hits
. . . Erwinia amylovora MR1	enterobacteria	1371	1	Erwinia amylovora MR1 hits
. . . Erwinia amylovora ATCC BAA-2158	enterobacteria	1369	1	Erwinia amylovora ATCC BAA-2158 hits
. . . Erwinia sp. Ejp617	enterobacteria	1347	2	Erwinia sp. Ejp617 hits
. . . Erwinia pyrifoliae	enterobacteria	1339	3	Erwinia pyrifoliae hits
. . . Erwinia pyrifoliae Ep1/96	enterobacteria	1339	1	Erwinia pyrifoliae Ep1/96 hits
. . . Erwinia pyrifoliae DSM 12163	enterobacteria	1339	1	Erwinia pyrifoliae DSM 12163 hits
. . . Erwinia piriflorinigrans	enterobacteria	1269	1	Erwinia piriflorinigrans hits
. . . Erwinia piriflorinigrans CFBP 5888	enterobacteria	1269	1	Erwinia piriflorinigrans CFBP 5888 hits
. . . Erwinia tasmaniensis	enterobacteria	1242	1	Erwinia tasmaniensis hits
. . . Erwinia tasmaniensis Et1/99	enterobacteria	1242	1	Erwinia tasmaniensis Et1/99 hits
. . . Erwinia psidii	enterobacteria	1211	2	Erwinia psidii hits
. . . Erwinia tracheiphila	enterobacteria	1184	3	Erwinia tracheiphila hits
. . . Erwinia tracheiphila PSU-1	enterobacteria	1183	1	Erwinia tracheiphila PSU-1 hits
. . . Erwinia mallotivora	enterobacteria	1163	2	Erwinia mallotivora hits
. . . Erwinia sp. AG740	enterobacteria	930	2	Erwinia sp. AG740 hits
. . Pantoea vagans	enterobacteria	1180	2	Pantoea vagans hits
. . Pantoea agglomerans	enterobacteria	1176	5	Pantoea agglomerans hits
. . Pantoea sp. VS1	enterobacteria	1160	2	Pantoea sp. VS1 hits
. . Pantoea sp. paga	enterobacteria	1152	2	Pantoea sp. paga hits
. . Pantoea stewartii	enterobacteria	1149	11	Pantoea stewartii hits

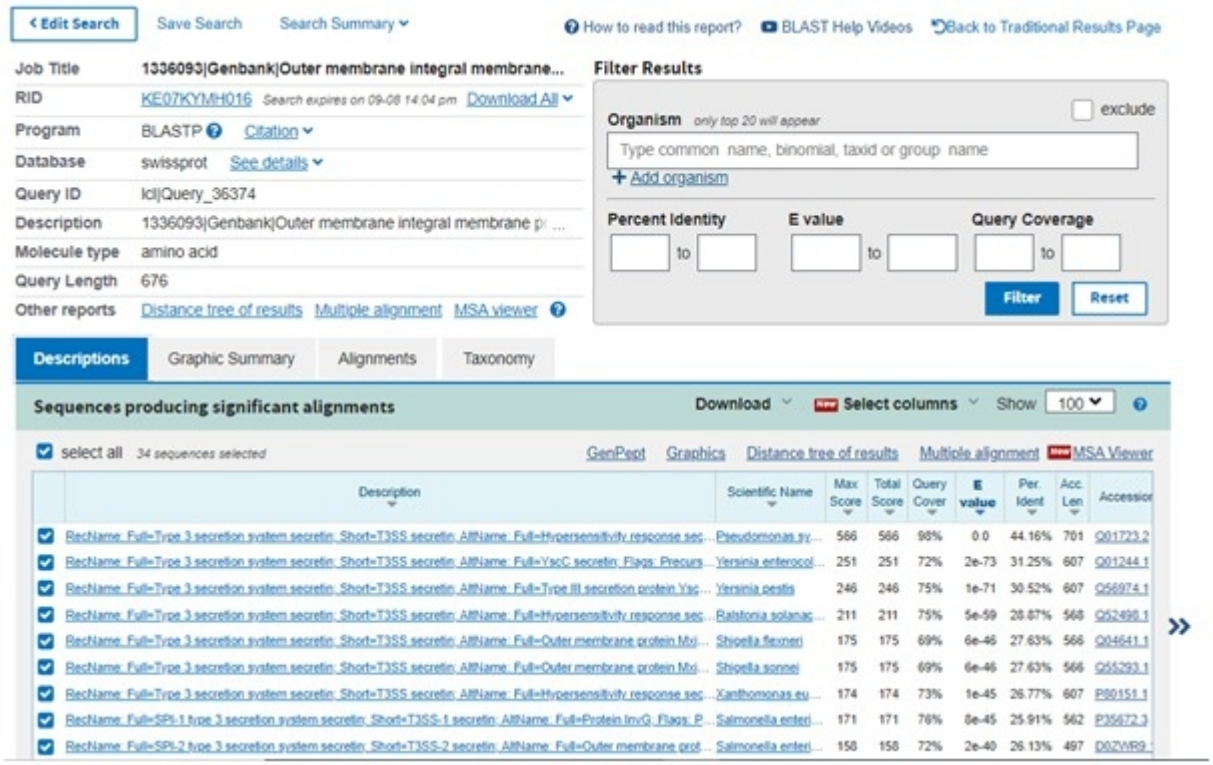
Descriptions | **Graphic Summary** | Alignments | Taxonomy

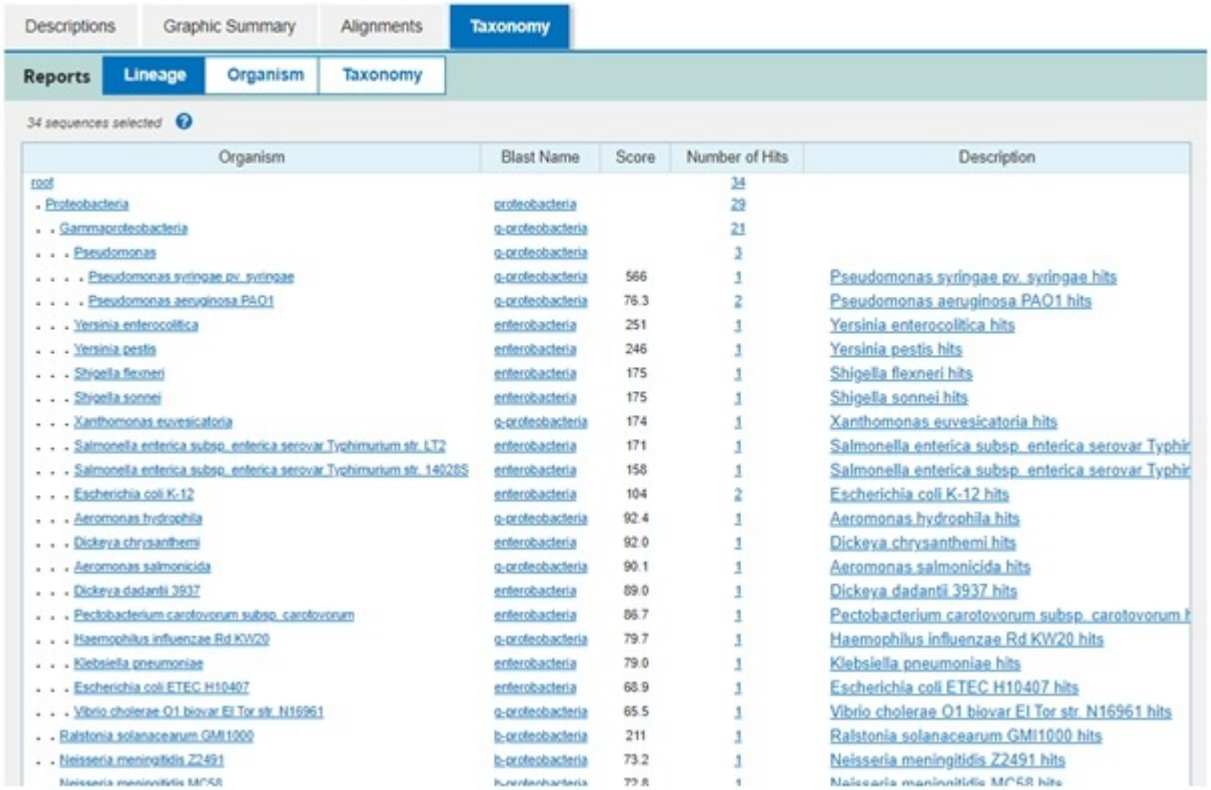
hover to see the title *click to show alignments* *Show Conserved Domains* Alignment Scores < 40 | 40 - 50 | 50 - 80 | 80 - 200 | >= 200

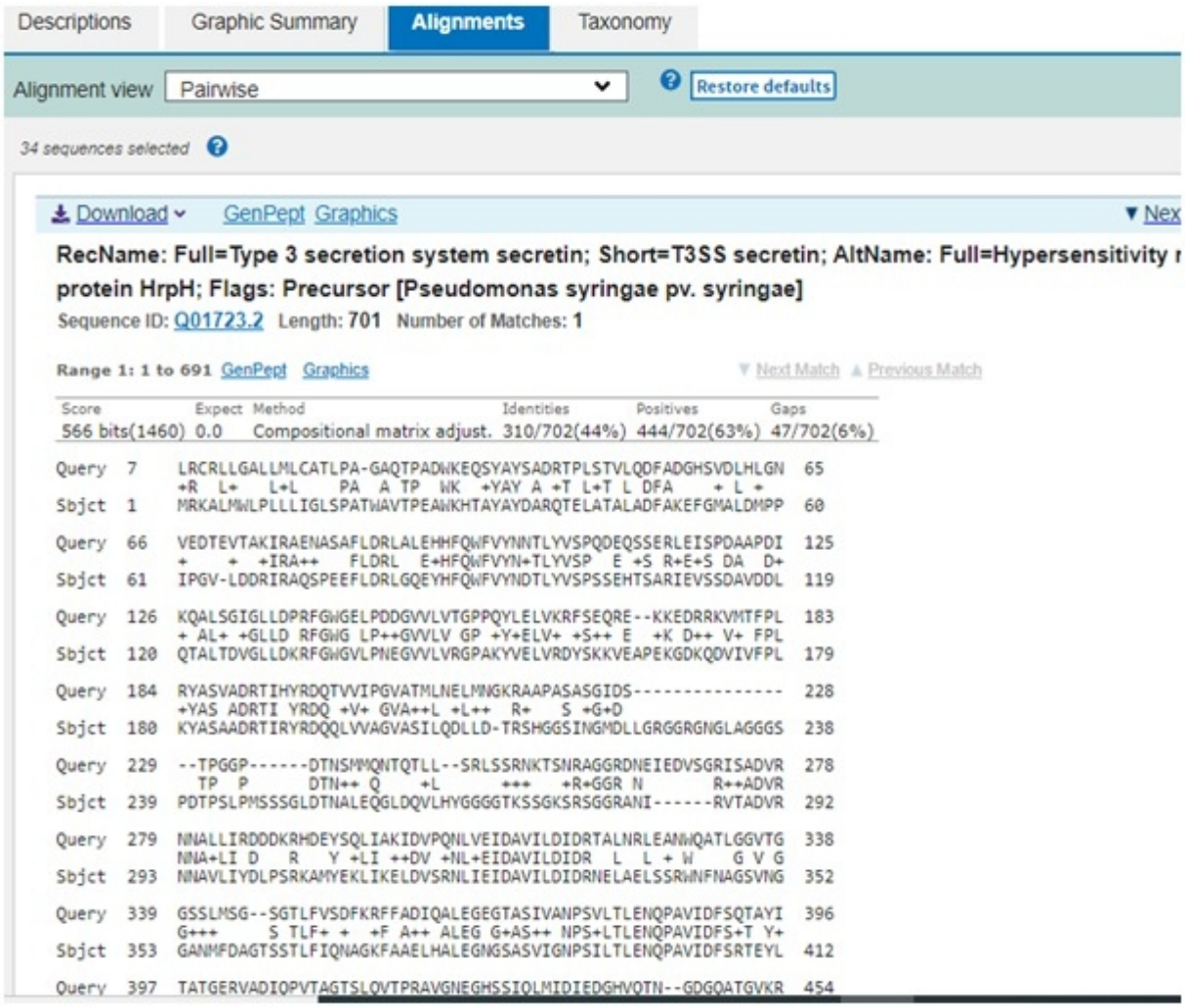
34 sequences selected

Putative conserved domains have been detected, click on the image below for detailed results.

Distribution of the top 34 Blast Hits on 34 subject sequences


Edit Search
Save Search
Search Summary
How to read this report?
BLAST Help Videos
Back to Traditional Results Page
Job Title 1336093|Genbank|Outer membrane integral membrane...
RID KE07KYMH016 Search expires on 09-08 14:04 pm Download All
Program BLASTP Citation
Database swissprot See details
Query ID lcl|Query_36374
Description 1336093|Genbank|Outer membrane integral membrane p ...
Molecule type amino acid
Query Length 676
Other reports Distance tree of results Multiple alignment MSA viewer
Filter Results
Organism only top 20 will appear
exclude
Type common name, binomial, taxid or group name
Add organism
Percent Identity
E value
Query Coverage
to
Filter
Reset
Descriptions
Graphic Summary
Alignments
Taxonomy
Sequences producing significant alignments
Download
Select columns
Show 100
select all 34 sequences selected
GenPept
Graphics
Distance tree of results
Multiple alignment
MSA Viewer
Description
Scientific Name
Max Score
Total Score
Query Cover
E value
Per. Ident
Acc. Len
Accession
RecName: Full=Type 3 secretion system secretin; Short=T3SS secretin; AltName: Full=Hypersensitivity response sec... Pseudomonas sy... 566 566 98% 0.0 44.16% 701 Q01723.2
RecName: Full=Type 3 secretion system secretin; Short=T3SS secretin; AltName: Full=YscC secretin; Flags: Precurs... Yersinia enterocol... 251 251 72% 2e-73 31.25% 607 Q01244.1
RecName: Full=Type 3 secretion system secretin; Short=T3SS secretin; AltName: Full=Type III secretion protein Ysc... Yersinia pestis 246 246 75% 1e-71 30.52% 607 Q56974.1
RecName: Full=Type 3 secretion system secretin; Short=T3SS secretin; AltName: Full=Hypersensitivity response sec... Ralstonia solanac... 211 211 75% 5e-59 28.87% 568 Q52498.1
RecName: Full=Type 3 secretion system secretin; Short=T3SS secretin; AltName: Full=Outer membrane protein MxI... Shigella flexneri 175 175 69% 6e-46 27.63% 566 Q04641.1
RecName: Full=Type 3 secretion system secretin; Short=T3SS secretin; AltName: Full=Outer membrane protein MxI... Shigella sonnei 175 175 69% 6e-46 27.63% 566 Q55293.1
RecName: Full=Type 3 secretion system secretin; Short=T3SS secretin; AltName: Full=Hypersensitivity response sec... Xanthomonas eu... 174 174 73% 1e-45 26.77% 607 P80151.1
RecName: Full=SPI-1 type 3 secretion system secretin; Short=T3SS-1 secretin; AltName: Full=Protein InvG; Flags: P... Salmonella enteri... 171 171 76% 8e-45 25.91% 562 P35672.3
RecName: Full=SPI-2 type 3 secretion system secretin; Short=T3SS-2 secretin; AltName: Full=Outer membrane prot... Salmonella enteri... 158 158 72% 2e-40 26.13% 497 D0ZWR9


Descriptions
Graphic Summary
Alignments
Taxonomy
Reports
Lineage
Organism
Taxonomy
34 sequences selected
Organism
Blast Name
Score
Number of Hits
Description
root 34
Proteobacteria proteobacteria 29
Gammaproteobacteria g-proteobacteria 21
Pseudomonas g-proteobacteria 3
Pseudomonas syringae pv. syringae g-proteobacteria 566 1 Pseudomonas syringae pv. syringae hits
Pseudomonas aeruginosa PAO1 g-proteobacteria 76.3 2 Pseudomonas aeruginosa PAO1 hits
Yersinia enterocolitica enterobacteria 251 1 Yersinia enterocolitica hits
Yersinia pestis enterobacteria 246 1 Yersinia pestis hits
Shigella flexneri enterobacteria 175 1 Shigella flexneri hits
Shigella sonnei enterobacteria 175 1 Shigella sonnei hits
Xanthomonas euvesicatoria g-proteobacteria 174 1 Xanthomonas euvesicatoria hits
Salmonella enterica subsp. enterica serovar Typhimurium str. LT2 enterobacteria 171 1 Salmonella enterica subsp. enterica serovar Typhi
Salmonella enterica subsp. enterica serovar Typhimurium str. 14028S enterobacteria 158 1 Salmonella enterica subsp. enterica serovar Typhi
Escherichia coli K-12 enterobacteria 104 2 Escherichia coli K-12 hits
Aeromonas hydrophila g-proteobacteria 92.4 1 Aeromonas hydrophila hits
Dickeya chrysanthemi enterobacteria 92.0 1 Dickeya chrysanthemi hits
Aeromonas salmonicida g-proteobacteria 90.1 1 Aeromonas salmonicida hits
Dickeya dadantii 3937 enterobacteria 89.0 1 Dickeya dadantii 3937 hits
Pectobacterium carotovorum subsp. carotovorum enterobacteria 86.7 1 Pectobacterium carotovorum subsp. carotovorum
Haemophilus influenzae Rd KW20 g-proteobacteria 79.7 1 Haemophilus influenzae Rd KW20 hits
Klebsiella pneumoniae enterobacteria 79.0 1 Klebsiella pneumoniae hits
Escherichia coli ETEC H10407 enterobacteria 68.9 1 Escherichia coli ETEC H10407 hits
Vibrio cholerae O1 biovar El Tor str. N16961 g-proteobacteria 65.5 1 Vibrio cholerae O1 biovar El Tor str. N16961 hits
Ralstonia solanacearum GMI1000 b-proteobacteria 211 1 Ralstonia solanacearum GMI1000 hits
Neisseria meningitidis Z2491 b-proteobacteria 73.2 1 Neisseria meningitidis Z2491 hits


Descriptions
Graphic Summary
Alignments
Taxonomy
Alignment view
Pairwise
Restore defaults
34 sequences selected
Download
GenPept Graphics
Next
RecName: Full=Type 3 secretion system secretin; Short=T3SS secretin; AltName: Full=Hypersensitivity r
protein HrpH; Flags: Precursor [Pseudomonas syringae pv. syringae]
Sequence ID: Q01723.2 Length: 701 Number of Matches: 1
Range 1: 1 to 691 GenPept Graphics
Next Match
Previous Match
Score Expect Method Identities Positives Gaps
566 bits(1460) 0.0 Compositional matrix adjust. 310/702(44%) 444/702(63%) 47/702(6%)
Query 7 LRCRLLGALLMLCATLPA-GAQTPADWKEQSYAYSADRTPLSTVLQDFADGHSVDLHLGN 65
+R L+ L+L PA A TP WK +YAY A +T L+T L DFA + L +
Sbjct 1 MRKALMWLPLLLIGLSPATWAVTPEAWKHTAYAYDARQTELATALADFAKEFGMALDMPP 60
Query 66 VEDTEVTAKIRAENASAFLDRLALEHHFQWFVYNNTLYVSPQDEQSSERLEISPDAAPDI 125
+ + +IRA++ FLDRL E+HFQWFVYN+TLYVSP E +S R+E+S DA D+
Sbjct 61 IPGV-LDDRIRAQSPEEFLDRLGQEYHFQWFVYNDTLYVSPSSEHTSARIEVSSDAVDDL 119
Query 126 KQALSGIGLLDPRFGWGELPDDGVVLVTGPPQYLELVKRFSEQRE--KKEDRRKVMTFPL 183
+ AL+ +GLLD RFGWG LP++GVVLV GP +Y+ELV+ +S++ E +K D++ V+ FPL
Sbjct 120 QTALTDVGLLDKRFGWGVLPNEGVVLVRGPAKYVELVRDYSKKVEAPEKGDKQDVIVFPL 179
Query 184 RYASVADRTIHYRDQTVVIPGVATMLNELMNGKRAAPASASGIDS---------------- 228
+YAS ADRTI YRDQ +V+ GVA++L +L++ R+ S +G+D
Sbjct 180 KYASAADRTIRYRDQQLVVAGVASILQDLLD-TRSHGGSINGMDLLGRGGRGNGLAGGGS 238
Query 229 --TPGGP------DTNSMMQNTQTLL--SRLSSRNKTSNRAGGRDNEIEDVSGRISADVR 278
TP P DTN++ Q +L +++ +R+GGR N R++ADVR
Sbjct 239 PDTPSLPMSSSGLDTNALEQGLDQVLHYGGGGTKSSGKSRSGGRANI------RVTADVR 292
Query 279 NMALLIRDDDKRHDEYSQLIAKIDVPQNLVEIDAVILDIDRTALNRLEANWQATLGGVTG 338
NMA+LI D R Y +LI ++DV +NL+EIDAVILDIDR L L + W G V G
Sbjct 293 NMAVLIYDLPSRKAMYEKLIKELDVSRNLIEIDAVILDIDRNELAELSSRWNFNAGSVNG 352
Query 339 GSSLMSG--SGTLFVSDFKRFFADIQALEGEGTASIVANPSVLTLENQPAVIDFSQTAYI 396
G+++ S TLF+ + +F A++ ALEG G+AS++ NPS+LTLENQPAVIDFS+T Y+
Sbjct 353 GANMFDAGTSSTLFIQNAGKFAAELHALEGNGSASVIGNPSILTLENQPAVIDFSRTEYL 412
Query 397 TATGERVADIQPVTAGTSLQVTPRAVGNEGHSSIQLMIDIEDGHVQTN--GDGQATGVKR 454

6. For this exercise, you will need to access Genbank by going to NCBI website and using the dropdown menu to search "Nucleotide" Note that the definition of the coding strand is the strand of DNA within the gene that is identical to the transcript (for genetic code use codon table). On the other hand, the template strand is a strand that is complementary to the coding strand.

1. Use the following Accession number to access the nucleotide sequence in the Genbank : CU329670

2. Go to the FEATURES section of the record.

3. Link to the CDS to gain access to the first 5662 nucleotides of the sequence.

4. Name the protein product of the CDS.

→ Protein product is RecQ type DNA helicase. It contains 1887 amino acids.

5. Write the first four amino acids (starting from the N terminus)

→The first four amino acids are methionine M ,valine V,valine V, alanine A.

6. Write the nucleotide sequence of the coding strand that corresponds to these amino acids.

→5'ATGGTCGTCGCT3' coding strand

7. Write the nucleotide sequence of the template strand that corresponds to these amino acids.

→3'TACCAGCAGCGA5' template strand

8. Using the sequence shown in the record, give the nucleotide number range that corresponds to these amino acids.