ACCESS BIOINFORMATICS DATABASES WITH BIO-PYTHON

This project is aimed to deploy python-based programming pipelines and scripts to automate biological data retrieval and analysis.

2. ENTREZ

This section fetches PUBMED literature and nucleotide sequences using ENTREZ. ENTREZ is NCBI's primary text search and retrieval system that integrates the PUBMED database of biomedical literature wtih 38 other literature and molecular databases, including DNA and protein sequences, structure, gene, genome, genetic variation and gene expression.

Import Modules

```
In [32]: from Bio import Entrez
```

Entrez requires users to put in their emails to run queries:

```
In [34]: Entrez.email = "jfxy0518@gmail.com"
```

I ran the code below to identify the types of databases which can be accessed using the Entrez module.

In [35]:
```
handle = Entrez.einfo()
record = Entrez.read(handle)
record["DbList"]
```

Out[35]:

```
['pubmed', 'protein', 'nuccore', 'ipg', 'nucleotide', 'structure', 'genome', 'annotinfo',
'assembly', 'bioproject', 'biosample', 'blastdbinfo', 'books', 'cdd', 'clinvar', 'gap', 'g
applus', 'grasp', 'dbvar', 'gene', 'gds', 'geoprofiles', 'homologene', 'medgen', 'mesh',
'nlmcatalog', 'omim', 'orgtrack', 'pmc', 'popset', 'proteinclusters', 'pcassay', 'protfa
m', 'pccompound', 'pcsubstance', 'seqannot', 'snp', 'sra', 'taxonomy', 'biocollections',
'gtr']
```

2.1. PUBMED

In this step, I attempt to fetch PubMed literature data.

In [36]:
```
handle = Entrez.einfo(db="pubmed")
record = Entrez.read(handle)
record["DbInfo"]["Description"]
```

Out[36]:

```
'PubMed bibliographic record'
```

As shown, PubMed is a bibliographic database containing bibliographic records...

```
In [37]: record["DbInfo"]["Count"]
```

Out[37]:

```
'35953109'
```

...with a count of 35937484 bibliographic records in this database (as of July 13th, 2023).

I then filtered the data of choice by using the e-search module of Entrez. This fetches all the literature or data containing the term 'biopython' in their title:

```
In [38]: handle = Entrez.esearch(db="pubmed", term="biopython")
         record = Entrez.read(handle)
record["IdList"]
```

Out[38]:

```
['36818783', '36245797', '36094101', '35497637', '35496474', '35402671', '34735950', '3448
4417', '34434786', '34189012', '33994075', '33902722', '33809815', '33242467', '32044951',
'31762715', '31278684', '31069053', '30013827', '29641230']
```

An alternative way to fetch literature details is to use the IdList instead of the term 'biopython'. To do that, I replaced the parameter 'esearch' with 'esummary', and pass in the IDs I got from the previous search. I formatted it into a specific type, showing the author, title, publication date and journal name using the *for* loop:

```
In [39]: handle = Entrez.esummary(db="pubmed", id='36818783,36245797')
         records = Entrez.parse(handle)


for record in records:
    print(record['AuthorList'],record['Title'],record['PubDate'],record['FullJournalName'])
```

```
['Olds CG', 'Berta-Thompson JW', 'Loucks JJ', 'Levy RA', 'Wilson AW'] Applying a modified
metabarcoding approach for the sequencing of macrofungal specimens from fungarium collecti
ons. 2023 Jan-Feb Applications in plant sciences
['Nallasamy V', 'Seshiah M'] Energy Profile Bayes and Thompson Optimized Convolutional Neu
ral Network protein structure prediction. 2023 Neural computing & applications
```

2.2. Nucleotide

In this section, I fetch nucleotide sequence records using the same Entrez module. Within the e-search parameter, I put 'nucleotide' as my database, and set it to retrieve 10 records. The ID list of all the nucleotide sequences which have the term 'severe acute respiratory syndrome' is found:

In [40]:
```
handle = Entrez.esearch(db="nucleotide",retmax=10, term="Severe acute respiratory syndrome")
record = Entrez.read(handle)
record["IdList"]
```

Out[40]:

```
['2542475402', '2542475384', '2542404305', '2542404289', '2542404272', '2542404258', '2542
404242', '2542404228', '2542404213', '2542404198']
```

Then, I fetched the record of these individual IDs. The retrieval type is set to 'gb' (Genbank), and the output is set to be in a text format:

```
In [41]: handle = Entrez.efetch(db="nucleotide", id="2531721439", rettype="gb", retmode="text")
         print(handle.read())
```

```
LOCUS       LC773238               29685 bp    RNA     linear   VRL 08-JUL-2023
DEFINITION  Severe acute respiratory syndrome coronavirus 2
            SARS-CoV-2/human/Japan/kmumc011145/2023 RNA, nearly complete
            genome.
ACCESSION   LC773238
VERSION     LC773238.1
DBLINK      BioProject: PRJDB16147
            BioSample: SAMD00627887
            Sequence Read Archive: DRR489579
KEYWORDS    .
SOURCE      Severe acute respiratory syndrome coronavirus 2
  ORGANISM  Severe acute respiratory syndrome coronavirus 2
            Viruses; Riboviria; Orthornavirae; Pisuviricota; Pisoniviricetes;
            Nidovirales; Cornidovirineae; Coronaviridae; Orthocoronavirinae;
            Betacoronavirus; Sarbecovirus; Severe acute respiratory
            syndrome-related coronavirus.
REFERENCE   1
  AUTHORS   Nakamori,Y. and Kashihara,M.
  TITLE     Clinical Experience of Treatment of Immunocompromised Individuals
            with Persistent SARS-CoV-2 Infection Based on Drug Resistance
            Mutations Determined by Genomic Analysis: A Descriptive Study
  JOURNAL   Unpublished
REFERENCE   2  (bases 1 to 29685)
  AUTHORS   Kashihara,M. and Inoue,A.
  TITLE     Direct Submission
```

```
  JOURNAL   Submitted (26-JUN-2023) Contact:Akira Inoue KANSAI MEDICAL
            UNIVERSITY MEDICAL CENTER; Fumizono-cho, Moriguchi-city, Osaka
            570-8507, Japan
COMMENT     ##Genome-Assembly-Data-START##
            Assembly Method       :: generateConsensus v. 5.16.0.12
            Genome Coverage       :: 14366x
            Sequencing Technology :: Genexus
            ##Genome-Assembly-Data-END##
FEATURES             Location/Qualifiers
     source          1..29685
                     /organism="Severe acute respiratory syndrome coronavirus
                     2"
                     /mol_type="genomic RNA"
                     /isolate="SARS-CoV-2/human/Japan/kmumc011145/2023"
                     /isolation_source="Nasopharyngeal swabs"
                     /host="Homo sapiens"
                     /db_xref="taxon:2697049"
                     /country="Japan"
                     /collection_date="2023-01-17"
     gene            224..21504
                     /gene="ORF1ab"
     CDS             join(224..13417,13417..21504)
                     /gene="ORF1ab"
                     /ribosomal_slippage
                     /codon_start=1
                     /product="ORF1ab polyprotein"
                     /protein_id="BEK81918.1"
                     /translation="MESLVPGFNEKTHVQLSLPVLQVRDVLVRGFGDSVEEVLSEARQ
                     HLKDGTCGLVEVEKGVLPQLEQPYVFIKRSDARTAPHGHVMVELVAELEGIQYGRSGE
                     TLGVLVPHVGEIPVAYRKVLLRKNGNKGAGGHRYGADLKSFDLGDELGTDPYEDFQEN
                     WNTKHSSGVTRELMRELNGGAYTRYVDNNFCGPDGYPLECIKDLLARAGKASCTLSEQ
                     LDFIDTKRGVYCCREHEHEIAWYTERSEKSYELQTPFEIKLAKKFDTFNGECPNFVFP
                     LNSIIKTIQPRVEKKKLDGFMGRIRSVYPVASPNECNQMCLSTLMKCDHCGETSWQTG
                     DFVKATCEFCGTENLTKEGATTCGYLPQNAVVKIYCPACHNSEVGPEHSLAEYHNESG
                     LKTILRKGGRTIAFGGCVFSYVGCHNKCAYWVPRASANIGCNHTGVVGEGSEGLNDNL
                     LEILQKEKVNINIVGDFKLNEEIAIILASFSASTSAFVETVKGLDYKAFKQIVESCGN
                     FKVTKGKAKKGAWNIGEQKSILSPLYAFASEAARVVRSIFSRTLETAQNSVRVLQKAA
```

ITILDGISQYSLRLIDAMMFTSDLATNNLVVMAYITGGVVQLTSQWLTNIFGTVYEKL
KPVLDWLEEKFKEGVEFLRDGWEIVKFISTCACEIVGGQIVTCAKEIKESVQTFFKLV
NKFLALCADSIIIGGAKLKALNLGETFVTHSKGLYRKCVKSREETGLLMPLKAPKEII
FLEGETLPTEVLTEEVVLKTGDLQPLEQPTSEAVEAPLVGTPVCINGLMLLEIKDTEK
YCALAPNMMVTNNTFTLKGGAPTKVTFGDDTVIEVQGYKSVNIIFELDERIDKVLNEK
CSAYTVELGTEVNEFACVVADAVIKTLQPVSELLTPLGIDLDEWSMATYYLFDESGEF
KLASHMYCSFYPPDEDEEEGDCEEEEFEPSTQYEYGTEDDYQGKPLEFGATSAALQPE
EEQEEDWLDDDSQQTVGQQDGSEDNQTTTIQTIVEVQPQLEMELTPVVQTIEVNSFSG
YLKLTDNVYIKNADIVEEAKKVKPTVVVNAANVYLKHGGGVAGALNKATNNAMQVESD
DYIATNGPLKVGGSCVLSGHNLAKHCLHVVGPNVNKGEDIQLLKSAYENFNQHEVLLA
PLLSAGIFGADPIHSLRVCVDTVRTNVYLAVFDKNLYDKLVSSFLEMKSEKQVEQKIA
EIPKEEVKPFITESKPSVEQRKQDDKKIKACVEEVTTTLEETKFLTENLLLYIDINGN
LHPDSATLVSDIDITFLKKDAPYIVGDVVQEGVLTAVVIPTKKASGTTEMLAKALRKV
PTDNYITTYPGQGLNGYTVEEAKTVLKKCKSAFYILPSIISNEKQEILGTVSWNLREM
LAHAEETRKLMPVCVETKAIVSTIQRKYKGIKIQEGVVDYGARFYFYTSKTTVASLIN
TLNDLNETLVTMPLGYVTHGLNLEEAARYMRSLKVPATVSVSSPDAVTAYNGYLTSSS
KTPEEHFIETISLAGSYKDWSYSGQSTQLGIEFLKRGDKSVYYTSNPTTFHLDGEVIT
FDNLKTLLSLREVRTIKVFTTVDNINLHTQVVDMSMTYGQQFGPTYLDGADVTKIKPH
NSHEGKTFYVLPNDDTLRVEAFEYYHTTDPSFLGRYMSALNHTKKWKYPQVNGLTSIK
WADNNCYLATALLTLQQIELKFNPPALQDAYYRARAGEAANFCALILAYCNKTVGELG
DVRETMSYLFQHANLDSCKRVLNVVCKTCGQQQTTLKGVEAVMYMGTLSYEQFKKGVQ
IPCTCGKQATKYLVQQESPFVMMSAPPAQYELKHGTFTCASEYTGNYQCGHYKHITSK
ETLYCIDGALLTKSSEYKGPITDVFYKENSYTTTIKPVTYKLDGVVCTEIDPKLDNYY
KKDNSYFTEQPIDLVPNQPYPNASFDNFKFVCDNIKFADDLNQLTGYKKPASRELKVT
FFPDLNGDVVAIDYKHYTPSFKKGAKLLHKPIVWHVNNATNKATYKPNTWCIRCLWST
KPVETSNSFDVLKSEDAQGMDNLACEDLKPVSEEVVENPTIQKDVLECNVKTTEVVGD
IILKPANNSLKITEEVGHTDLMAAYVDNSSLTIKKPNELSRVLGLKILATHGLAAVNS
VPWDTIANYAKPFLNKVVSTTTNIVTRCLNRVCTNYMPYFFTLLLQLCTFTRSTNSRI
KASMPTTIAKNTVKSVGKFCLEASFNYLKSPNFSKLINIIIWFLLLSVCLGSLIYSTA
ALGVLMSNLGMPSYCTGYREGYLNSTNVTIATYCTGSIPCSVCLSGLDSLDTYPSLET
IQITISSFKWDLTAFGLVAEWFLAYILFTRFFYVLGLAAIMQLFFSYFAVHFISNSWL
MWLIINLVQMAPISAMVRMYIFFASFYYVWKSYVHVVDGCNSSTCMMCYKRNRATRVE
CTTIVNGVRRSFYVYANGGKGFCKLHNWNCVNCDTFCAGSTFISDEVARDLSLQFKRP
INPTDQSSYIVDSVTVKNGSIHLYFDKAGQKTYERHSLSHFVNLDNLRANNTKGSLPI
NVIVFDGKSKCEESSAKSASVYYSQLMCQPILLLDQALVSDVGDSAEVAVKMFDAYVN
TFSSTFNVPMEKLKTLVATAEAELAKNVSLDNVLSTFISAARQGFVDSDVETKDVVEC
LKLSHQSDIEVTGDSCNNYMLTYNKVENMTPRDLGACIDCSARHINAQVAKSHNIALI

WNVKDFMSLSEQLRKQIRSAAKKNNLPFKLTCATTRQVVNVVTTKIALKGGKIVNNWL
KQLIKVTLVFLFVAAIFYLITPVHVMSKHTDFSSEIIGYKAIDGGVTRDIASTDTCFA
NKHADFDTWFSQRGGSYTNDKACPLIAAVITREVGFVVPGLPGTVLRTTNGDFLHFLP
RVFSAVGNICYTPSKLIEYTDFATSACVLAAECTIFKDASGKPVPYCYDTNVLEGSVA
YESLRPDTRYVLMDGSIIQFPNTYLEGSVRVVTTFDSEYCRHGTCERSEAGVCVSTSG
RWVLNNDYYRSLPGVFCGVDAVNLFTNMFTPLIQPIGALDISASIVAGGIVAIVVTCL
AYYFMRFRRAXXXXXXXXXXXXXXXXXXXXXXXXXXXFLPGVYSVIYLYLTFYLTN
DVSFLAHIQWMVMFTPLVPFWITIAYIICISTKHFYWFFSNYLKRRVVFNGVSFSTFE
EAALCTFLLNKEMYLKLRSDVLLPLTQYNRYLALYNKYKYFSGAMDTTSYREAACCHL
AKALNDFSNSGSDVLYQPPQISITSAVLQSGFRKMAFPSGKVEGCMVQVTCGTTTLNG
LWLDDVVYCPRHVICTSEDMLNPNYEDLLIRKSNHNFLVQAGNVQLRVIGHSMQNCVL
KLKVDTANPKTPKYKFVRIQPGQTFSVLACYNGSPSGVYQCAMRHNFTIKGSFLNGSC
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
STLEQYVFCTVNALPETTADIVVFDEISMATNYDLSVVNARLCAKHYVYIGDPAQLPA
PRTLLTKGTLEPEYFNSVCRLMKTIGPDMFLGTCRRCPAEIVDTVSALVYDNKLKAHK
DKSAQCFKMFYKGVITHDVSSAINRPQIGVVREFLTRNPAWRKAVFISPYNSQNAVAS
KILGLPTQTVDSSQGSEYDYVIFTQTTETAHSCNVNRFNVAITRAKVGILCIMSDRDL
YDKLQFTSLEIPRRNVATLQAENVTGLFKDCSKVITGLHPTQAPTHLSVDTKFKTEGL
CVDVPGIPKDMTYRRLISMMGFKMNYQVNGYPNMFITREEAIRHVRAWIGFDVEGCHA
TREAVGTNLPLQLGFSTGVNLVAVPTGYVDTPNNTDFSRVSAKPPPGDQFKHLIPLMY
KGLPWNVVRIKIVQMLSDTLKNLSDRVVFVLWAHGFELTSMKYFVKIGPERTCCLCDR
RATCFSTASDTYACWHHSIGFDYVYNPFMIDVQQWGFTGNLQSNHDLYCQVHGNAHVA
SCDAIMTRCLAVHECFVKRVDWTIEYPIIGDELKINAACRKVQHMVVKAALLADKFPV
LHDIGNPKAIKCVPQADVEWKFYDAQPCSDKAYKIEELFYSYATHSDKFTDGVCLFWN
CNVDRYPANSIVCRFDTRVLSNLNLPGCDGGSLYVNKHAFHTPAFDKSAFVNLKQLPF
FYYSDSPCESHGKQVVSDIDYVPLKSATCITRCNLGGAVCRHHANEYRLYLDAYNMMI
SAGFSLWVYKQFDTYNLWNTFTRLQSLENVAFNVVNKGHFDGQQGEVPVSIINNTVYT
KVDGVDVELFENKTTLPVNVAFELWAKRNIKPVPEVKILNNLGVDIAANTVIWDYKRD
APAHISTIGVCSMTDIAKKPIETICAPLTVFFDGRVDGQVDLFRNARNGVLITEGSVK
GLQPSVGPKQASLNGVTLIGEAVKTQFNYYKKVDGVVQQLPETYFTQSRNLQEFKPRS
QMEIDFLELAMDEFIERYKLEGYAFEHIVYGDFSHSQLGGLHLLIGLAKRFKESPFEL
EDFIPMDSTVKNYFITDAQTGSSKCVCSVIDLLLDDFVEIIKSQDLSVVSKVVKVTID
YTEISFMLWCKDGHVETFYPKLQSSQAWQPGVAMPNLYKMQRMLLEKCDLQNYGDSAT
LPKGIMMNVAKYTQLCQYLNTLTLAVPYNMRVIHFGAGSDKGVAPGTAVLRQWLPTGT
LLVDSDLNDFVSDADSTLIGDCATVHTANKWDLIISDMYDPKTKNVTKENDSKEGFFT
YICGFIQQKLALGGSVAIKITEHSWSADLYKLMGHFAWWTAFVTNVNASSSEAFLIGC

```
                     NYLGKPREQIDGYVMHANYIFWRNTNPIQLSSYSLFDMSKFPLKLRGTAVMSLKEGQI
                     NDMILSLLSKGRLIIRENNRVVISSDVLVNN"
     mat_peptide     16186..17988
                     /gene="ORF1ab"
                     /product="helicase"
     mat_peptide     17989..19569
                     /gene="ORF1ab"
                     /product="3'-to-5' exonuclease"
     mat_peptide     19570..20607
                     /gene="ORF1ab"
                     /product="endoRNAse"
     mat_peptide     20608..21501
                     /gene="ORF1ab"
                     /product="2'-O-ribose methyltransferase"
     CDS             224..13432
                     /gene="ORF1ab"
                     /codon_start=1
                     /product="ORF1a polyprotein"
                     /protein_id="BEK81919.1"
                     /translation="MESLVPGFNEKTHVQLSLPVLQVRDVLVRGFGDSVEEVLSEARQ
                     HLKDGTCGLVEVEKGVLPQLEQPYVFIKRSDARTAPHGHVMVELVAELEGIQYGRSGE
                     TLGVLVPHVGEIPVAYRKVLLRKNGNKGAGGHRYGADLKSFDLGDELGTDPYEDFQEN
                     WNTKHSSGVTRELMRELNGGAYTRYVDNNFCGPDGYPLECIKDLLARAGKASCTLSEQ
                     LDFIDTKRGVYCCREHEHEIAWYTERSEKSYELQTPFEIKLAKKFDTFNGECPNFVFP
                     LNSIIKTIQPRVEKKKLDGFMGRIRSVYPVASPNECNQMCLSTLMKCDHCGETSWQTG
                     DFVKATCEFCGTENLTKEGATTCGYLPQNAVVKIYCPACHNSEVGPEHSLAEYHNESG
                     LKTILRKGGRTIAFGGCVFSYVGCHNKCAYWVPRASANIGCNHTGVVGEGSEGLNDNL
                     LEILQKEKVNINIVGDFKLNEEIAIILASFSASTSAFVETVKGLDYKAFKQIVESCGN
                     FKVTKGKAKKGAWNIGEQKSILSPLYAFASEAARVVRSIFSRTLETAQNSVRVLQKAA
                     ITILDGISQYSLRLIDAMMFTSDLATNNLVVMAYITGGVVQLTSQWLTNIFGTVYEKL
                     KPVLDWLEEKFKEGVEFLRDGWEIVKFISTCACEIVGGQIVTCAKEIKESVQTFFKLV
                     NKFLALCADSIIIGGAKLKALNLGETFVTHSKGLYRKCVKSREETGLLMPLKAPKEII
                     FLEGETLPTEVLTEEVVLKTGDLQPLEQPTSEAVEAPLVGTPVCINGLMLLEIKDTEK
                     YCALAPNMMVTNNTFTLKGGAPTKVTFGDDTVIEVQGYKSVNIIFELDERIDKVLNEK
                     CSAYTVELGTEVNEFACVVADAVIKTLQPVSELLTPLGIDLDEWSMATYYLFDESGEF
                     KLASHMYCSFYPPDEDEEEGDCEEEEFEPSTQYEYGTEDDYQGKPLEFGATSAALQPE
                     EEQEEDWLDDDSQQTVGQQDGSEDNQTTTIQTIVEVQPQLEMELTPVVQTIEVNSFSG
```

YLKLTDNVYIKNADIVEEAKKVKPTVVVNAANVYLKHGGGVAGALNKATNNAMQVESD
DYIATNGPLKVGGSCVLSGHNLAKHCLHVVGPNVNKGEDIQLLKSAYENFNQHEVLLA
PLLSAGIFGADPIHSLRVCVDTVRTNVYLAVFDKNLYDKLVSSFLEMKSEKQVEQKIA
EIPKEEVKPFITESKPSVEQRKQDDKKIKACVEEVTTTLEETKFLTENLLLYIDINGN
LHPDSATLVSDIDITFLKKDAPYIVGDVVQEGVLTAVVIPTKKASGTTEMLAKALRKV
PTDNYITTYPGQGLNGYTVEEAKTVLKKCKSAFYILPSIISNEKQEILGTVSWNLREM
LAHAEETRKLMPVCVETKAIVSTIQRKYKGIKIQEGVVDYGARFYFYTSKTTVASLIN
TLNDLNETLVTMPLGYVTHGLNLEEAARYMRSLKVPATVSVSSPDAVTAYNGYLTSSS
KTPEEHFIETISLAGSYKDWSYSGQSTQLGIEFLKRGDKSVYYTSNPTTFHLDGEVIT
FDNLKTLLSLREVRTIKVFTTVDNINLHTQVVDMSMTYGQQFGPTYLDGADVTKIKPH
NSHEGKTFYVLPNDDTLRVEAFEYYHTTDPSFLGRYMSALNHTKKWKYPQVNGLTSIK
WADNNCYLATALLTLQQIELKFNPPALQDAYYRARAGEAANFCALILAYCNKTVGELG
DVRETMSYLFQHANLDSCKRVLNVVCKTCGQQQTTLKGVEAVMYMGTLSYEQFKKGVQ
IPCTCGKQATKYLVQQESPFVMMSAPPAQYELKHGTFTCASEYTGNYQCGHYKHITSK
ETLYCIDGALLTKSSEYKGPITDVFYKENSYTTTIKPVTYKLDGVVCTEIDPKLDNYY
KKDNSYFTEQPIDLVPNQPYPNASFDNFKFVCDNIKFADDLNQLTGYKKPASRELKVT
FFPDLNGDVVAIDYKHYTPSFKKGAKLLHKPIVWHVNNATNKATYKPNTWCIRCLWST
KPVETSNSFDVLKSEDAQGMDNLACEDLKPVSEEVVENPTIQKDVLECNVKTTEVVGD
IILKPANNSLKITEEVGHTDLMAAYVDNSSLTIKKPNELSRVLGLKILATHGLAAVNS
VPWDTIANYAKPFLNKVVSTTTNIVTRCLNRVCTNYMPYFFTLLLQLCTFTRSTNSRI
KASMPTTIAKNTVKSVGKFCLEASFNYLKSPNFSKLINIIIWFLLLSVCLGSLIYSTA
ALGVLMSNLGMPSYCTGYREGYLNSTNVTIATYCTGSIPCSVCLSGLDSLDTYPSLET
IQITISSFKWDLTAFGLVAEWFLAYILFTRFFYVLGLAAIMQLFFSYFAVHFISNSWL
MWLIINLVQMAPISAMVRMYIFFASFYYVWKSYVHVVDGCNSSTCMMCYKRNRATRVE
CTTIVNGVRRSFYVYANGGKGFCKLHNWNCVNCDTFCAGSTFISDEVARDLSLQFKRP
INPTDQSSYIVDSVTVKNGSIHLYFDKAGQKTYERHSLSHFVNLDNLRANNTKGSLPI
NVIVFDGKSKCEESSAKSASVYYSQLMCQPILLLDQALVSDVGDSAEVAVKMFDAYVN
TFSSTFNVPMEKLKTLVATAEAELAKNVSLDNVLSTFISAARQGFVDSDVETKDVVEC
LKLSHQSDIEVTGDSCNNYMLTYNKVENMTPRDLGACIDCSARHINAQVAKSHNIALI
WNVKDFMSLSEQLRKQIRSAAKKNNLPFKLTCATTRQVVNVVTTKIALKGGKIVNNWL
KQLIKVTLVFLFVAAIFYLITPVHVMSKHTDFSSEIIGYKAIDGGVTRDIASTDTCFA
NKHADFDTWFSQRGGSYTNDKACPLIAAVITREVGFVVPGLPGTVLRTTNGDFLHFLP
RVFSAVGNICYTPSKLIEYTDFATSACVLAAECTIFKDASGKPVPYCYDTNVLEGSVA
YESLRPDTRYVLMDGSIIQFPNTYLEGSVRVVTTFDSEYCRHGTCERSEAGVCVSTSG
RWVLNNDYYRSLPGVFCGVDAVNLFTNMFTPLIQPIGALDISASIVAGGIVAIVVTCL
AYYFMRFRRAXXXXXXXXXXXXXXXXXXXXXXXXXXXXFLPGVYSVIYLYLTFYLTN
DVSFLAHIQWMVMFTPLVPFWITIAYIICISTKHFYWFFSNYLKRRVVFNGVSFSTFE

```
                     EAALCTFLLNKEMYLKLRSDVLLPLTQYNRYLALYNKYKYFSGAMDTTSYREAACCHL
                     AKALNDFSNSGSDVLYQPPQISITSAVLQSGFRKMAFPSGKVEGCMVQVTCGTTTLNG
                     LWLDDVVYCPRHVICTSEDMLNPNYEDLLIRKSNHNFLVQAGNVQLRVIGHSMQNCVL
                     KLKVDTANPKTPKYKFVRIQPGQTFSVLACYNGSPSGVYQCAMRHNFTIKGSFLNGSC
                     GSVGFNIDYDCVSFCYMHHMELPTGVHAGTDLEGNFYGPFVDRQTAQAAGTDTTITVN
                     VLAWLYAAVINGDRWFLNRFTTTLNDFNLVAMKYNYEPLTQDHVDILGPLSAQTGIAV
                     LDMCASLKELLQNGMNGRTILGSALLEDEFTPFDVVRQCSGVTFQSAVKRTIKGTHHW
                     LLLTILTSLLVLVQSTQWSLFFFLYENAFLPFAMGIIAMSAFAMMFVKHKHAFLCLFL
                     LPSLATVAYFNMVYMPASWVMRIMTWLDMVDTSLKLKDCVMYASAVVLLILMTARTVY
                     DDGARRVWTLMNVLTLVYKVYYGNALDQAISMWALIISVTSNYSGVVTTVMFLARGIV
                     FMCVEYCPIFFITGNTLQCIMLVYCFLGYFCTCYFGLFCLLNRYFRLTLGVYDYLVST
                     QEFRYMNSQGLLPPKNSIDAFKLNIKLLGVGGKPCIKVATVQSKMSDVKCTSVVLLSV
                     LQQLRVESSSKLWAQCVQLHNDILLAKDTTEAFEKMVSLLSVLLSMQGAVDINKLCEE
                     MLDNRATLQAIASEFSSLPSYAAFATAQEAYEQAVANGDSEVVLKKLKKSLNVAKSEF
                     DRDAAMQRKLEKMADQAMTQMYKQARSEDKRAKVTSAMQTMLFTMLRKLDNDALNNII
                     NNARDGCVPLNIIPLTTAAKLMVVIPDYNTYKNTCDGTTFTYASALWEIQQVVDADSK
                     IVQLSEISMDNSPNLAWPLIVTALRANSAVKLQNNELSPVALRQMSCAAGTTQTACTD
                     DNALAYYNTTKGGRFVLALLSDLQDLKWARFPKSDGTGTIYTELEPPCRFVTDTPKGP
                     KVKYLYFIKGLNNLNRGMVLGSLAATVRLQAGNATEVPANSTVLSFCAFAVDAAKAYK
                     DYLASGGQPITNCVKMLCTHTGTGQAITVTPEANMDQESFGGASCCLYCRCHIDHPNP
                     KGFCDLKGKYVQIPTTCANDPVGFTLKNTVCTVCGMWKGYGCSCDQLREPMLQSADAQ
                     SFLNGFAV"
mat_peptide          224..763
                     /gene="ORF1ab"
                     /product="leader protein"
mat_peptide          764..2677
                     /gene="ORF1ab"
                     /product="nsp2"
mat_peptide          2678..8512
                     /gene="ORF1ab"
                     /product="nsp3"
mat_peptide          8513..10012
                     /gene="ORF1ab"
                     /product="nsp4"
mat_peptide          10013..10930
                     /gene="ORF1ab"
                     /product="3C-like proteinase"
```

```
mat_peptide     10931..11791
                /gene="ORF1ab"
                /product="nsp6"
mat_peptide     11792..12040
                /gene="ORF1ab"
                /product="nsp7"
mat_peptide     12041..12634
                /gene="ORF1ab"
                /product="nsp8"
mat_peptide     12635..12973
                /gene="ORF1ab"
                /product="nsp9"
mat_peptide     12974..13390
                /gene="ORF1ab"
                /product="nsp10"
mat_peptide     13391..13429
                /gene="ORF1ab"
                /product="nsp11"
assembly_gap    9435..9520
                /estimated_length=86
                /gap_type="within scaffold"
                /linkage_evidence="align genus"
mat_peptide     join(13391..13417,13417..16185)
                /gene="ORF1ab"
                /product="RNA-dependent RNA polymerase"
stem_loop       13425..13452
                /gene="ORF1ab"
                /note="Coronavirus frameshifting stimulation element
                stem-loop 1"
stem_loop       13437..13491
                /gene="ORF1ab"
                /note="Coronavirus frameshifting stimulation element
                stem-loop 2"
gene            21512..25318
                /gene="S"
CDS             21512..25318
                /gene="S"
```

```
                     /codon_start=1
                     /product="surface glycoprotein"
                     /protein_id="BEK81920.1"
                     /translation="MFVFLVLLPLVSSQCVNLITRTQSYTNSFTRGVYYPDKVFRSSV
                     LHSTQDLFLPFFSNVTWFHAISGTNGTKRFDNPVLPFNDGVYFASTEKSNIIRGWIFG
                     TTLDSKTQSLLIVNNATNVVIKVCEFQFCNDPFLDVYYHKNNKSWMESEFRVYSSANN
                     CTFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGYFKIYSKHTPINLGRDLPQGFSAL
                     EPLVDLPIGINITRFQTLLALHRSYLTPGDSSSGWTAGAAAYYVGYLQPRTFLLKYNE
                     NGTITDAVDCALDPLSETKCTLKSFTVEKGIYQTSNFRVQPTESIVRFPNITNLCPFD
                     EVFNATRFASVYAWNRKRISNCVADYSVLYNFAPFFAFKCYGVSPTKLNDLCFTNVYA
                     DSFVIRGNEVSQIAPGQTGNIADYNYKLPDDFTGCVIAWNSNKLDSKVGGNYNYRYRL
                     FRKSNLKPFERDISTEIYQAGNKPCNGVAGVNCYFPLQSYGFRPTYGVGHQPYRVVVL
                     SFELLHAPATVCGPKKSTNLVKNKCVNFNFNGLTGTGVLTESNKKFLPFQQFGRDIAD
                     TTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLYQGVNCTEVPVAIHADQL
                     TPTWRVYSTGSNVFQTRAGCLIGAEYVNNSYECDIPIGAGICASYQTQTKSHRRARSV
                     ASQSIIAYTMSLGAENSVAYSNNSIAIPTNFTISVTTEILPVSMIKTSVDCTMYICGD
                     STECSNLLLQYGSFCTQLKRALTGIAVEQDKNTQEVFAQVKQIYKTPPIKYFGGFNFS
                     QILPDPSKPSKRSFIEDLLFNKVTLADAGFIKQYGDCLGDIAARDLICAQKFNGLTVL
                     PPLLTDEMIAQYTSALLAGTITSGWTFGAGAALQIPFAMQMAYRFNGIGVTQNVLYEN
                     QKLIANQFNSAIGKIQDSLSSTASALGKLQDVVNHNAQALNTLVKQLSSKFGAISSVL
                     NDILSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRSSANLAATKMSECVLG
                     QSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPAICHDGKAHFPREG
                     VFVSNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPLQPELDSFKEE
                     LDKYFKNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDLQELGKYEQY
                     IKWPWYIWLGFIAGLIAIVMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSEPVLKG
                     VKLHYT"
gene                 25327..26154
                     /gene="ORF3a"
CDS                  25327..26154
                     /gene="ORF3a"
                     /codon_start=1
                     /product="ORF3a protein"
                     /protein_id="BEK81921.1"
                     /translation="MDLFMRIFTIGTVTLKQGEIKDATPSDFVRATATIPIQASLPFG
                     WLIVGVALLAVFQSASKIITLKKRWQLALSKGVHFVCNLLLLFVTVYSHLLLVAAGLE
                     APFLYLYALVYFLQSINFVRIIMRLWLCWKCRSKNPLLYDANYFLCWHTNCYDYCIPY
                     NSVTSSIVITSGDGTTSPISEHDYQIGGYTEKWESGVKDCVVLHSYFTSDYYQLYSTQ
```

```
                     LSTDIGVEHVTFFIYNKIVDEPEEHVQIHTIDGSSGVVNPVMEPIYDEPTTTTSVPL"
gene                 26179..26406
                     /gene="E"
CDS                  26179..26406
                     /gene="E"
                     /codon_start=1
                     /product="envelope protein"
                     /protein_id="BEK81922.1"
                     /translation="MYSFVSEEIGTLIVNSVLLFLAFVVFLLVTLAILTALRLCAYCC
                     NIVNVSLVKPSFYVYSRVKNLNSSRVPDLLV"
gene                 26457..27125
                     /gene="M"
CDS                  26457..27125
                     /gene="M"
                     /codon_start=1
                     /product="membrane glycoprotein"
                     /protein_id="BEK81923.1"
                     /translation="MANSNGTITVEELKKLLEEWNLVIGFLFLTWICLLQFAYANRNR
                     FLYIIKLIFLWLLWPVTLTCFVLAAVYRINWITGGIAIAMACLVGLMWLSYFIASFRL
                     FARTRSMWSFNPETNILLNVPLHGTILTRPLLESELVIGAVILRGHLRIAGHHLGRCD
                     IKDLPKEITVATSRTLSYYKLGASQRVAGDSGFAAYSRYRIGNYKLNTDHSSSSDNIA
                     LLVQ"
gene                 27136..27321
                     /gene="ORF6"
CDS                  27136..27321
                     /gene="ORF6"
                     /codon_start=1
                     /product="ORF6 protein"
                     /protein_id="BEK81924.1"
                     /translation="MFHLVDFQVTIAEILLIIMRTFKVSIWNLDYIINLIIKNLSKSL
                     TENKYSQLDEEQPMEID"
gene                 27328..27693
                     /gene="ORF7a"
CDS                  27328..27693
                     /gene="ORF7a"
                     /codon_start=1
                     /product="ORF7a protein"
```

```
                     /protein_id="BEK81925.1"
                     /translation="MKIILFLALITLATCELYHYQECVRGTTVLLKEPCSSGTYEGNS
                     PFYPLADNKFALTCFSTQFAFACPDGVKHVYQLRARSVSPKLFIRQEEVQELYSPIFL
                     IVAAIVFITLCFTLKRKTE"
gene                 27690..27821
                     /gene="ORF7b"
CDS                  27690..27821
                     /gene="ORF7b"
                     /codon_start=1
                     /product="ORF7b"
                     /protein_id="BEK81926.1"
                     /translation="MIELSLIDFYLCFLAFLLFLVLIMLIIFWFSLELQDHNETCHA"
gene                 27828..28193
                     /gene="ORF8"
CDS                  27828..28193
                     /gene="ORF8"
                     /codon_start=1
                     /product="ORF8 protein"
                     /protein_id="BEK81927.1"
                     /translation="MKFLVFLGXXXXXXXXXXXXXXQSCTQHQPYVVDDPCPIHFYSK
                     WYIRVGARKSAPLIELCVDEAGSKSPIQYIDIGNYTVSCLLFTINCQEPKLGSLVVRC
                     SFYEDFLEYHDVRVVLDFI"
assembly_gap         27853..27892
                     /estimated_length=40
                     /gap_type="within scaffold"
                     /linkage_evidence="align genus"
gene                 28208..29458
                     /gene="N"
CDS                  28208..29458
                     /gene="N"
                     /codon_start=1
                     /product="nucleocapsid phosphoprotein"
                     /protein_id="BEK81928.1"
                     /translation="MSDNGPQNQRNALRITFGGPSDSTGSNQNGGARSKQRRPQGLPN
                     NTASWFTALTQHGKEDLKFPRGQGVPINTNSSPDDQIGYYRRATRRIRGGDGKMKDLS
                     PRWYFYYLGTGPEAGLPYGANKDGIIWVATEGALNTPKDHIGTRNPANNAAIVLQLPQ
                     GTTLPKGFYAEGSRGGSQASSRSSSRSRNSSRNSTPGSSKRTSPARMAGNGGDAALAL
```

```
                     LLLDRLNQLESKMSGKGQQQQGQTVTKKSAAEASKKPRQKRTATKAYNVTQAFGRRGP
                     EQTQGNFGDQELIRQGTDYKHWPQIAQFAPSASAFFGMSRIGMEVTPSGTWLTYTGAI
                     KLDDKDPNFKDQVILLNKHIDAYKTFPPTEPKKDKKKADETQALPQRQKKQQTVTLL
                     PAADLDDFSKQLQQSMSRADSTQA"
     assembly_gap    29469..29502
                     /estimated_length=34
                     /gap_type="within scaffold"
                     /linkage_evidence="align genus"
     gene            <29503..29599
                     /gene="ORF10"
     CDS             <29503..29599
                     /gene="ORF10"
                     /note="no start codon"
                     /codon_start=2
                     /product="ORF10 protein"
                     /protein_id="BEK81929.1"
                     /translation="AFPFTIYSLLLCRMNSRNYIAQVDVVNFNLT"
     stem_loop       29534..29569
                     /gene="ORF10"
                     /note="Coronavirus 3' UTR pseudoknot stem-loop 1"
     stem_loop       29554..29582
                     /gene="ORF10"
                     /note="Coronavirus 3' UTR pseudoknot stem-loop 2"
     stem_loop       29653..29667
                     /note="Coronavirus 3' stem-loop II-like motif (s2m)"
ORIGIN
        1 ttgatctctt gtagatctgt tctctaaacg aactttaaaa tctgtgtggc tgtcactcgg
       61 ctgcatgctt agtgcactca cgcagtataa ttaataacta attactgtcg ttgacaggac
      121 acgagtaact cgtctatctt ctgcaggctg cttacggttt cgtccgtgtt gcagccgatc
      181 atcagcacat ctaggttttg tccgggtgtg accgaaaggt aagatggaga gccttgtccc
      241 tggtttcaac gagaaaacac acgtccaact cagtttgcct gttttacagg ttcgcgacgt
      301 gctcgtacgt ggctttggag actccgtgga ggaggtctta tcagaggcac gtcaacatct
      361 taaagatggc acttgtggct tagtagaagt tgaaaaaggc gttttgcctc aacttgaaca
      421 gccctatgtg ttcatcaaac gttcggatgc tcgaactgca cctcatggtc atgttatggt
      481 tgagctggta gcagaactcg aaggcattca gtacggtcgt agtggtgaga cacttggtgt
      541 ccttgtccct catgtgggcg aaataccagt ggcttaccgc aaggttcttc ttcgtaagaa
      601 cggtaataaa ggagctggtg gccataggta cggcgccgat ctaaagtcat ttgacttagg
```

```
 661 cgacgagctt ggcactgatc cttatgaaga ttttcaagaa aactggaaca ctaaacatag
 721 cagtggtgtt acccgtgaac tcatgcgtga gcttaacgga ggggcataca ctcgctatgt
 781 cgataacaac ttctgtggcc ctgatggcta ccctcttgag tgcattaaag accttctagc
 841 acgtgctggt aaagcttcat gcactttgtc cgaacaactg gactttattg acactaagag
 901 gggtgtatac tgctgccgtg aacatgagca tgaaattgct tggtacacgg aacgttctga
 961 aaagagctat gaattgcaga cacctttga aattaaattg gcaaagaaat ttgacacctt
1021 caatggggaa tgtccaaatt ttgtatttcc cttaaattcc ataatcaaga ctattcaacc
1081 aagggttgaa aagaaaaagc ttgatggctt tatgggtaga attcgatctg tctatccagt
1141 tgcgtcacca aatgaatgca accaaatgtg cctttcaact ctcatgaagt gtgatcattg
1201 tggtgaaact tcatggcaga cgggcgattt tgttaaagcc acttgcgaat tttgtggcac
1261 tgagaatttg actaaagaag gtgccactac ttgtggttac ttaccccaaa atgctgttgt
1321 taaaatttat tgtccagcat gtcacaattc agaagtagga cctgagcata gtcttgccga
1381 ataccataat gaatctggct tgaaaaccat tcttcgtaag ggtggtcgca ctattgcctt
1441 tggaggctgt gtgttctctt atgttggttg ccataacaag tgtgcctatt gggttccacg
1501 tgctagcgct aacataggtt gtaaccatac aggtgttgtt ggagaaggtt ccgaaggtct
1561 taatgacaac cttcttgaaa tacttcaaaa agagaaagtc aacatcaata ttgttggtga
1621 ctttaaactt aatgaagaga tcgccattat tttggcatct ttttctgctt ccacaagtgc
1681 ttttgtggaa actgtgaaag gtttggatta taaagcattc aaacaaattg ttgaatcctg
1741 tggtaatttt aaagttacaa aaggaaaagc taaaaaaggt gcctggaata ttggtgaaca
1801 gaaatcaata ctgagtcctc tttatgcatt tgcatcagag gctgctcgtg ttgtacgatc
1861 aattttctcc cgcactcttg aaactgctca aaattctgtg cgtgttttac agaaggccgc
1921 tataacaata ctagatggaa tttcacagta ttcactgaga ctcattgatg ctatgatgtt
1981 cacatctgat ttggctacta acaatctagt tgtaatggcc tacattacag gtggtgttgt
2041 tcagttgact tcgcagtggc taactaacat ctttggcact gtttatgaaa aactcaaacc
2101 cgtccttgat tggcttgaag agaagtttaa ggaaggtgta gagtttctta gagacggttg
2161 ggaaattgtt aaatttatct caacctgtgc ttgtgaaatt gtcggtggac aaattgtcac
2221 ctgtgcaaag gaaattaagg agagtgttca gacattcttt aagcttgtaa ataaattttt
2281 ggctttgtgt gctgactcta tcattattgg tggagctaaa cttaaagcct tgaatttagg
2341 tgaaacattt gtcacgcact caaagggatt gtacagaaag tgtgttaaat ccagagaaga
2401 aactggccta ctcatgcctc taaaagcccc aaaagaaatt atcttcttag agggagaaac
2461 acttcccaca gaagtgttaa cagaggaagt tgtcttgaaa actggtgatt tacaaccatt
2521 agaacaacct actagtgaag ctgttgaagc tccattggtt ggtacaccag tttgtattaa
2581 cgggcttatg ttgctcgaaa tcaaagacac agaaaagtac tgtgcccttg cacctaatat
2641 gatggtaaca aacaatacct tcacactcaa aggcggtgca ccaacaaagg ttacttttgg
2701 tgatgacact gtgatagaag tgcaaggtta caagagtgtg aatatcattt ttgaacttga
2761 tgaaaggatt gataaagtac ttaatgagaa gtgctctgcc tatacagttg aactcggtac
2821 agaagtaaat gagttcgcct gtgttgtggc agatgctgtc ataaaaactt tgcaaccagt
```

2881 atctgaatta cttacaccac tgggcattga tttagatgag tggagtatgg ctacatacta
2941 cttatttgat gagtctggtg agtttaaatt ggcttcacat atgtattgtt ctttttaccc
3001 tccagatgag gatgaagaag aaggtgattg tgaagaagaa gagtttgagc catcaactca
3061 atatgagtat ggtactgaag atgattacca aggtaaacct ttggaatttg gtgccacttc
3121 tgctgctctt caacctgaag aagagcaaga agaagattgg ttagatgatg atagtcaaca
3181 aactgttggt caacaagacg gcagtgagga caatcagaca actactattc aaacaattgt
3241 tgaggttcaa cctcaattag agatggaact tacaccagtt gttcagacta ttgaagtgaa
3301 tagttttagt ggttatttaa aacttactga caatgtatac attaaaaatg cagacattgt
3361 ggaagaagct aaaaaggtaa aaccaacagt ggttgttaat gcagccaatg tttaccttaa
3421 acatggagga ggtgttgcag gagccttaaa taaggctact aacaatgcca tgcaagttga
3481 atctgatgat tacatagcta ctaatggacc acttaaagtg ggtggtagtt gtgttttaag
3541 cggacacaat cttgctaaac actgtcttca tgttgtcggc ccaaatgtta acaaaggtga
3601 agacattcaa cttcttaaga gtgcttatga aaattttaat cagcacgaag ttctacttgc
3661 accattatta tcagctggta tttttggtgc tgaccctata cattctttaa gagtttgtgt
3721 agatactgtt cgcacaaatg tctacttagc tgtctttgat aaaaatctct atgacaaact
3781 tgtttcaagc tttttggaaa tgaagagtga aaagcaagtt gaacaaaaga tcgctgagat
3841 tcctaaagag gaagttaagc catttataac tgaaagtaaa ccttcagttg aacagagaaa
3901 acaagatgat aagaaaatca aagcttgtgt tgaagaagtt acaacaactc tggaagaaac
3961 taagttcctc acagaaaact tgttacttta tattgacatt aatggcaatc ttcatccaga
4021 ttctgccact cttgttagtg acattgacat cactttctta aagaaagatg ctccatatat
4081 agtgggtgat gttgttcaag agggtgtttt aactgctgtg gttataccta ctaaaaaggc
4141 tagtggcact actgaaatgc tagcgaaagc tttgagaaaa gtgccaacag acaattatat
4201 aaccacttac ccgggtcagg gtttaaatgg ttacactgta gaggaggcaa agacagtgct
4261 taaaaagtgt aaaagtgctt tttacattct accatctatt atctctaatg agaagcaaga
4321 aattcttgga actgtttctt ggaatttgcg agaaatgctt gcacatgcag aagaaacacg
4381 caaattaatg cctgtctgtg tggaaactaa agccatagtt tcaactatac agcgtaaata
4441 taagggtatt aaaatacaag agggtgtggt tgattatggt gctagatttt acttttacac
4501 cagtaaaaca actgtagcgt cacttatcaa cacacttaac gatctaaatg aaactcttgt
4561 tacaatgcca cttggctatg taacacatgg cttaaatttg gaagaagctg ctcggtatat
4621 gagatctctc aaagtgccag ctacagtttc tgtttcttca cctgatgctg ttacagcgta
4681 taatggttat cttacttctt cttctaaaac acctgaagaa cattttattg aaaccatctc
4741 acttgctggt tcctataaag attggtccta ttctggacaa tctacacaac taggtataga
4801 atttcttaag agaggtgata aaagtgtata ttacactagt aatcctacca cattccacct
4861 agatggtgaa gttatcacct ttgacaatct taagacactt ctttctttga gagaagtgag
4921 gactattaag gtgtttacaa cagtagacaa cattaacctc cacacgcaag ttgtggacat
4981 gtcaatgaca tatggacaac agtttggtcc aacttatttg gatggagctg atgttactaa
5041 aataaaacct cataattcac atgaaggtaa aacattttat gttttaccta atgatgacac

5101 tctacgtgtt gaggcttttg agtactacca cacaactgat cctagttttc tgggtaggta
5161 catgtcagca ttaaatcaca ctaaaaagtg gaaataccca caagttaatg gtttaacttc
5221 tattaaatgg gcagataaca actgttatct tgccactgca ttgttaacac tccaacaaat
5281 agagttgaag tttaatccac ctgctctaca agatgcttat tacagagcaa gggctggtga
5341 agctgctaac ttttgtgcac ttatcttagc ctactgtaat aagacagtag gtgagttagg
5401 tgatgttaga gaaacaatga gttacttgtt tcaacatgcc aatttagatt cttgcaaaag
5461 agtcttgaac gtggtgtgta aaacttgtgg acaacagcag acaaccctta agggtgtaga
5521 agctgttatg tacatgggca cactttctta tgaacaattt aagaaaggtg ttcagatacc
5581 ttgtacgtgt ggtaaacaag ctacaaaata tctagtacaa caggagtcac cttttgttat
5641 gatgtcagca ccacctgctc agtatgaact taagcatggt acatttactt gtgctagtga
5701 gtacactggt aattaccagt gtggtcacta taaacatata acttctaaag aaactttgta
5761 ttgcatagac ggtgctttac ttacaaagtc ctcagaatac aaaggtccta ttacggatgt
5821 tttctacaaa gaaaacagtt acacaacaac cataaaacca gttacttata aattggatgg
5881 tgttgtttgt acagaaattg accctaagtt ggacaattat tataagaaag acaattctta
5941 tttcacagag caaccaattg atcttgtacc aaaccaacca tatccaaacg caagcttcga
6001 taattttaag tttgtatgtg ataatatcaa atttgctgat gatttaaacc agttaactgg
6061 ttataagaaa cctgcttcaa gagagcttaa agttacattt ttccctgact taaatggtga
6121 tgtggtggct attgattata aacactacac accctctttt aagaaaggag ctaaattgtt
6181 acataaacct attgtttggc atgttaacaa tgcaactaat aaagccacgt ataaaccaaa
6241 tacctggtgt atacgttgtc tttggagcac aaaaccagtt gaaacatcaa attcgtttga
6301 tgtactgaag tcagaggacg cgcagggaat ggataatctt gcctgcgaag atctaaaacc
6361 agtctctgaa gaagtagtgg aaaatcctac catacagaaa gacgttcttg agtgtaatgt
6421 gaaaactacc gaagttgtag gagacattat acttaaacca gcaaataata gtttaaaaat
6481 tacagaagag gttggccaca cagatctaat ggctgcttat gtagacaatt ctagtcttac
6541 tattaagaaa cctaatgaat tatctagagt attaggtttg aaaatccttg ctactcatgg
6601 tttagctgct gttaatagtg tcccttggga tactatagct aattatgcta agccttttct
6661 taacaaagtt gttagtacaa ctactaacat agttacacgg tgtttaaacc gtgtttgtac
6721 taattatatg ccttatttct ttactttatt gctacaattg tgtactttta ctagaagtac
6781 aaattctaga attaaagcat ctatgccgac tactatagca aagaatactg ttaagagtgt
6841 tggtaaattt tgtctagagg cttcatttaa ttatttgaag tcacctaatt tttctaaact
6901 gataaatatt ataatttggt ttttactatt aagtgtgtgc ctaggttctt taatctactc
6961 aaccgctgct ttaggtgttt taatgtctaa tttaggcatg ccttcttact gtactggtta
7021 cagagaaggc tatttgaact ctactaatgt cactattgca acctactgta ctggttctat
7081 accttgtagt gtttgtctta gtggtttaga ttctttagac acctatcctt ctttagaaac
7141 tatacaaatt accatttcat cttttaaatg ggatttaact gcttttggct tagttgcaga
7201 gtggtttttg gcatatattc ttttcactag gtttttctat gtacttggat tggctgcaat
7261 catgcaattg tttttcagct attttgcagt acattttatt agtaattctt ggcttatgtg

7321 gttaataatt aatcttgtac aaatggcccc gatttcagct atggttagaa tgtacatctt
7381 ctttgcatca ttttattatg tatggaaaag ttatgtgcat gttgtagacg gttgtaattc
7441 atcaacttgt atgatgtgtt acaaacgtaa tagagcaaca agagtcgaat gtacaactat
7501 tgttaatggt gttagaaggt ccttttatgt ctatgctaat ggaggtaaag gcttttgcaa
7561 actacacaat tggaattgtg ttaattgtga tacattctgt gctggtagta catttattag
7621 tgatgaagtt gcgagagact tgtcactaca gtttaaaaga ccaataaatc ctactgacca
7681 gtcttcttac atcgttgata gtgttacagt gaagaatggt tccatccatc tttactttga
7741 taaagctggt caaaagactt atgaaagaca ttctctctct cattttgtta acttagacaa
7801 cctgagagct aataacacta aaggttcatt gcctattaat gttatagttt ttgatggtaa
7861 atcaaaatgt gaagaatcat ctgcaaaatc agcgtctgtt tactacagtc agcttatgtg
7921 tcaacctata ctgttactag atcaggcatt agtgtctgat gttggtgata gtgcggaagt
7981 tgcagttaaa atgtttgatg cttacgttaa tacgttttca tcaacttta acgtaccaat
8041 ggaaaaactc aaaacactag ttgcaactgc agaagctgaa cttgcaaaga atgtgtcctt
8101 agacaatgtc ttatctactt ttatttcagc agctcggcaa gggtttgttg attcagatgt
8161 agaaactaaa gatgttgttg aatgtcttaa attgtcacat caatctgaca tagaagttac
8221 tggcgatagt tgtaataact atatgctcac ctataacaaa gttgaaaaca tgacaccccg
8281 tgaccttggt gcttgtattg actgtagtgc gcgtcatatt aatgcgcagg tagcaaaaag
8341 tcacaacatt gctttgatat ggaacgttaa agatttcatg tcattgtctg aacaactacg
8401 aaaacaaata cgtagtgctg ctaaaaagaa taacttacct tttaagttga catgtgcaac
8461 tactagacaa gttgttaatg ttgtaacaac aaagatagca cttaagggtg gtaaaattgt
8521 taataattgg ttgaagcagt taattaaagt tacacttgtg ttccttttg ttgctgctat
8581 tttctattta ataacacctg ttcatgtcat gtctaaacat actgactttt caagtgaaat
8641 cataggatac aaggctattg atggtggtgt cactcgtgac atagcatcta cagatacttg
8701 ttttgctaac aaacatgctg attttgacac atggtttagc cagcgtggtg gtagttatac
8761 taatgacaaa gcttgcccat tgattgctgc agtcataaca agagaagtgg gttttgtcgt
8821 gcctggtttg cctggcacgg tattacgcac aactaatggt gacttttgc atttcttacc
8881 tagagttttt agtgcagttg gtaacatctg ttacacacca tcaaaactta tagagtacac
8941 tgactttgca acatcagctt gtgttttggc tgctgaatgt acaatttta aagatgcttc
9001 tggtaagcca gtaccatatt gttatgatac caatgtacta gaaggttctg ttgcttatga
9061 aagtttacgc cctgacacac gttatgtgct catggatggc tctattattc aatttcctaa
9121 cacctacctt gaaggttctg ttagagtggt aacaactttt gattctgagt actgtaggca
9181 cggcacttgt gaaagatcag aagctggtgt ttgtgtatct actagtggta gatgggtact
9241 taacaatgat tattacagat ctttaccagg agttttctgt ggtgtagatg ctgtaaattt
9301 atttactaat atgtttacac cactaattca acctattggt gctttggaca tatcagcatc
9361 tatagtagct ggtggtattg tggctatcgt agtaacatgc cttgcctact attttatgag
9421 gtttagaaga gcttnnnnnn nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn
9481 nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn ttcttacctg gtgtttattc

```
 9541 tgttatttac ttgtacttga cattttatct tactaatgat gtttcttttt tagcacatat
 9601 tcagtggatg gttatgttca cacctttagt acctttctgg ataacaattg cttatatcat
 9661 ttgtatttcc acaaagcatt tctattggtt ctttagtaat tacctaaaga gacgtgtagt
 9721 ctttaatggt gtttccttta gtacttttga agaagctgcg ctgtgcacct ttttgttaaa
 9781 taaagaaatg tatctaaagt tgcgtagtga tgtgctatta cctcttacgc aatataatag
 9841 atacttagct ctttataata agtacaagta ttttagtgga gcaatggata caactagcta
 9901 cagagaagct gcttgttgtc atcttgcaaa ggctctcaat gacttcagta actcaggttc
 9961 tgatgttctt taccaaccac cacaaatctc tatcacctca gctgttttgc agagtggttt
10021 tagaaaaatg gcattcccat ctggtaaagt tgagggttgt atggtacaag taacttgtgg
10081 tacaactaca cttaacggtc tttggcttga tgacgtagtt tactgtccaa gacatgtgat
10141 ctgcacctct gaagatatgc ttaaccctaa ttatgaagat ttactcattc gtaagtctaa
10201 tcataatttc ttggtacagg ctggtaatgt tcaactcagg gttattggac attctatgca
10261 aaattgtgta cttaagctta aggttgatac agccaatcct aagacaccta agtataagtt
10321 tgttcgcatt caaccaggac agacttttc agtgttagct tgttacaatg gttcaccatc
10381 tggtgtttac caatgtgcta tgagacacaa tttcactatt aagggttcat tccttaatgg
10441 ttcatgtggt agtgttggtt ttaacataga ttatgactgt gtctcttttt gttacatgca
10501 ccatatggaa ttaccaactg gagttcatgc tggcacagac ttagaaggta acttttatgg
10561 acctttttgtt gacaggcaaa cagcacaagc agctggtacg gacacaacta ttacagttaa
10621 tgttttagct tggttgtacg ctgctgttat aaatggagac aggtggtttc tcaatcgatt
10681 taccacaact cttaatgact ttaaccttgt ggctatgaag tacaattatg aacctctaac
10741 acaagaccat gttgacatac taggacctct ttctgctcaa actggaattg ccgttttaga
10801 tatgtgtgct tcattaaaag aattactgca aaatggtatg aatggacgta ccatattggg
10861 tagtgcttta ttagaagatg aatttacacc ttttgatgtt gttagacaat gctcaggtgt
10921 tactttccaa agtgcagtga aaagaacaat caagggtaca caccactggt tgttactcac
10981 aattttgact tcacttttag ttttagtcca gagtactcaa tggtctttgt tctttttttt
11041 gtatgaaaat gccttttac cttttgctat gggtattatt gctatgtctg cttttgcaat
11101 gatgtttgtc aaacataagc atgcatttct ctgtttgttt ttgttacctt ctcttgccac
11161 tgtagcttat tttaatatgg tctatatgcc tgctagttgg gtgatgcgta ttatgacatg
11221 gttggatatg gttgatacta gtttgaagct aaaagactgt gttatgtatg catcagctgt
11281 agtgttacta atccttatga cagcaagaac tgtgtatgat gatggtgcta ggagagtgtg
11341 gacacttatg aatgtcttga cactcgttta taaagtttat tatggtaatg ctttagatca
11401 agccatttcc atgtgggctc ttataatctc tgttacttct aactactcag gtgtagttac
11461 aactgtcatg tttttggcca gaggtattgt ttttatgtgt gttgagtatt gccctatttt
11521 cttcataact ggtaatacac ttcagtgtat aatgctagtt tattgtttct taggctattt
11581 ttgtacttgt tactttggcc tcttttgttt actcaaccgc tactttagac tgactcttgg
11641 tgtttatgat tacttagttt ctacacagga gtttagatat atgaattcac agggactact
11701 cccacccaag aatagcatag atgccttcaa actcaacatt aaattgttgg gtgttggtgg
```

11761 caaaccttgt atcaaagtag ccactgtaca gtctaaaatg tcagatgtaa agtgcacatc
11821 agtagtctta ctctcagttt tgcaacaact cagagtagaa tcatcatcta aattgtgggc
11881 tcaatgtgtc cagttacaca atgacattct cttagctaaa gatactactg aagcctttga
11941 aaaaatggtt tcactacttt ctgttttgct ttccatgcag ggtgctgtag acataaacaa
12001 gctttgtgaa gaaatgctgg acaacagggc aaccttacaa gctatagcct cagagtttag
12061 ttcccttcca tcatatgcag cttttgctac tgctcaagaa gcttatgaac aggctgttgc
12121 taatggtgat tctgaagttg ttcttaaaaa gttgaagaag tctttgaatg tggctaaatc
12181 tgaatttgac cgtgatgcag ccatgcaacg taagttggaa aagatggctg atcaagctat
12241 gacccaaatg tataaacagg ctagatctga ggacaagagg gcaaaagtta ctagtgctat
12301 gcagacaatg cttttcacta tgcttagaaa gttggataat gatgcactca acaacattat
12361 caacaatgca agagatggtt gtgttccctt gaacataata cctcttacaa cagcagccaa
12421 actaatggtt gttataccag actataacac atataaaaat acgtgtgatg gtacaacatt
12481 tacttatgca tcagcattgt gggaaatcca acaggttgta gatgcagata gtaaaattgt
12541 tcaacttagt gaaattagta tggacaattc acctaattta gcatggcctc ttattgtaac
12601 agctttaagg gccaattctg ctgtcaaatt acagaataat gagcttagtc ctgttgcact
12661 acgacagatg tcttgtgctg ccggtactac acaaactgct tgcactgatg acaatgcgtt
12721 agcttactac aacacaacaa agggaggtag gtttgtactt gcactgttat ccgatttaca
12781 ggatttgaaa tgggctagat tccctaagag tgatggaact ggtactattt atacagaact
12841 ggaaccacct tgtaggtttg ttacagacac acctaaaggt cctaaagtga agtatttata
12901 ctttattaaa ggattaaaca acctaaatag aggtatggta cttggtagtt tagctgccac
12961 agtacgtcta caagctggta atgcaacaga agtgcctgcc aattcaactg tattatcttt
13021 ctgtgctttt gctgtagatg ctgctaaagc ttacaaagat tatctagcta gtgggggaca
13081 accaatcact aattgtgtta agatgttgtg tacacacact ggtactggtc aggcaataac
13141 agttacaccg gaagccaata tggatcaaga atcctttggt ggtgcatcgt gttgtctgta
13201 ctgccgttgc cacatagatc atccaaatcc taaaggattt tgtgacttaa aaggtaagta
13261 tgtacaaata cctacaactt gtgctaatga ccctgtgggt tttacactta aaaacacagt
13321 ctgtaccgtc tgcggtatgt ggaaaggtta tggctgtagt tgtgatcaac tccgcgaacc
13381 catgcttcag tcagctgatg cacaatcgtt tttaaacggg tttgcggtgt aagtgcagcc
13441 cgtcttacac cgtgcggcac aggcactagt actgatgtcg tatacagggc ttttgacatc
13501 tacaatgata aagtagctgg ttttgctaaa ttcctaaaaa ctaattgttg tcgcttccaa
13561 gaaaaggacg aagatgacaa tttaattgat tcttactttg tagttaagag acacactttc
13621 tctaactacc aacatgaaga aacaatttat aatttactta aggattgtcc agctgttgct
13681 aaacatgact tctttaagtt tagaatagac ggtgacatgg taccacatat atcacgtcaa
13741 cgtcttacta aatacacaat ggcagacctc gtctatgctt taaggcattt tgatgaaggt
13801 aattgtgaca cattaaaaga aatacttgtc acatacaatt gttgtgatga tgattatttc
13861 aataaaaagg actggtatga ttttgtagaa aacccagata tattacgcgt atacgccaac
13921 ttaggtgaac gtgtacgcca agctttgtta aaaacagtac aattctgtga tgccatgcga

13981 aatgctggta ttgttggtgt actgacatta gataatcaag atctcaatgg taactggtat
14041 gatttcggtg atttcataca aaccacgcca ggtagtggag ttcctgttgt agattcttat
14101 tattcattgt taatgcctat attaaccttg accagggctt taactgcaga gtcacatgtt
14161 gacactgact taacaaagcc ttacattaag tgggatttgt taaaatatga cttcacggaa
14221 gagaggttaa aactctttga ccgttatttt aaatattggg atcagacata ccacccaaat
14281 tgtgttaact gtttggatga cagatgcatt ctgcattgtg caaactttaa tgttttattc
14341 tctacagtgt tcccacttac aagttttgga ccactagtga gaaaaatatt tgttgatggt
14401 gttccatttg tagtttcaac tggataccac ttcagagagc taggtgttgt acataatcag
14461 gatgtaaact tacatagctc tagacttagt tttaaggaat tacttgtgta tgctgctgac
14521 cctgctatgc acgctgcttc tggtaatcta ttactagata aacgcactac gtgcttttca
14581 gtagctgcac ttactaacaa tgttgctttt caaactgtca aacccggtaa ttttaacaaa
14641 gacttctatg actttgctgt gtctaagggt ttctttaagg aaggaagttc tgttgaatta
14701 aaacacttct tctttgctca ggatggtaat gctgctatca gcgattatga ctactatcgt
14761 tataatctac caacaatgtg tgatatcaga caactactat ttgtagttga agttgttgat
14821 aagtactttg attgttacga tggtggctgt attaatgcta accaagtcat cgtcaacaac
14881 ctagacaaat cagctggttt tccatttaat aaatggggta aggctagact ttattatgat
14941 tcaatgagtt atgaggatca agatgcactt ttcgcatata caaaacgtaa tgtcatccct
15001 actataactc aaatgaatct taagtatgcc attagtgcaa agaatagagc tcgcaccgta
15061 gctggtgtct ctatctgtag tactatgacc aatagacagt ttcatcaaaa attattgaaa
15121 tcaatagccg ccactagagg agctactgta gtaattggaa caagcaaatt ctatggtggt
15181 tggcacaaca tgttaaaaac tgtttatagt gatgtagaaa accctcacct tatgggttgg
15241 gattatccta aatgtgatag agccatgcct aacatgctta gaattatggc ctcacttgtt
15301 cttgctcgca aacatacaac gtgttgtagc ttgtcacacc gtttctatag attagctaat
15361 gagtgtgctc aagtattgag tgaaatggtc atgtgtggcg gttcactata tgttaaacca
15421 ggtggaacct catcaggaga tgccacaact gcttatgcta atagtgtttt taacatttgt
15481 caagctgtca cggccaatgt taatgcactt ttatctactg atggtaacaa aattgccgat
15541 aagtatgtcc gcaatttaca acacagactt tatgagtgtc tctatagaaa tagagatgtt
15601 gacacagact ttgtgaatga gttttacgca tatttgcgta aacatttctc aatgatgata
15661 ctttctgatg atgctgttgt gtgttttaat agcacttatg catctcaagg tctagtggct
15721 agcataaaga actttaagtc agttctttat tatcaaaaca atgtttttat gtctgaagca
15781 aaatgttgga ctgagactga ccttactaaa ggacctcatg aattttgctc tcaacataca
15841 atgctagtta aacagggtga tgattatgtg taccttcctt acccagatcc atcaagaatc
15901 ctaggggctg gctgttttgt agatgatatc gtaaaaacag atggtacact tatgattgaa
15961 cggttcgtgt ctttagctat agatgcttac ccacttacta aacatcctaa tcaggagtat
16021 gctgatgtct ttcatttgta cttacaatac ataagaaagc tacatgatga gttaacagga
16081 cacatgttag acatgtattc tgttatgctt actaatgata acacttcaag gtattgggaa
16141 cctgagtttt atgaggctat gtacacaccg catacagtct tacaggctgt tggggcttgt

16201 gttctttgca attcacagac ttcattaaga tgtggtgctt gcatacgtag accattctta
16261 tgttgtaaat gctgttacga ccatgtcata tcaacatcac ataaattagt cttgtctgtt
16321 aatccgtatg tttgcaatgc tccaggttgt gatgtcacag atgtgactca actttactta
16381 ggaggtatga gctattattg taaatcacat aaaccaccca ttagttttcc attgtgtgct
16441 aatggacaag tttttggttt atataaaaat acatgtgttg gtagcgataa tgttactgac
16501 tttaatgcaa ttgcaacatg tgactggaca aatgctggtg attacatttt agctaacacc
16561 tgtactgaaa gactcaagct ttttgcagca gaaacgctca aagctactga ggagacattt
16621 aaactgtctt atggtattgc tactgtacgt gaagtgctgt ctgacagaga attacatctt
16681 tcatgggaag ttggtaaacc tagaccacca cttaaccgaa attatgtctt tactggttat
16741 cgtgtaacta aaaacagtaa agtacaaata ggagagtaca cctttgaaaa aggtgactat
16801 ggtgatgctg ttgtttaccg aggtacaaca acttacaaat taaatgttgg tgattatttt
16861 gtgctgacat cacatacagt aatgccatta agtgcaccta cattagtgcc acaagagcac
16921 tatgttagaa ttactggctt atacccaaca ctcaatatct cagatgagtt ttctagcaat
16981 gttgcaaatt atcaaaaggt tggtatgcaa aagtattcta cactccaggg accacctggt
17041 actggtaaga gtcattttgc tattggccta gctctctact acccttctgc tcgcatagtg
17101 tatacagctt gctctcatgc cgctgttgat gcactatgtg agaaggcatt aaaatatttg
17161 cctatagata aatgtagtag aattatacct gcacgtgctc gtgtagagtg ttttgataaa
17221 ttcaaagtga attcaacatt agaacagtat gtcttttgta ctgtaaatgc attgcctgag
17281 acgacagcag atatagttgt ctttgatgaa atttcaatgg ccacaaatta tgatttgagt
17341 gttgtcaatg ccagattatg tgctaagcac tatgtgtaca ttggcgaccc tgctcaatta
17401 cctgcaccac gcacattgct aactaagggc acactagaac cagaatattt caattcagtg
17461 tgtagactta tgaaaactat aggtccagac atgttcctcg gaacttgtcg gcgttgtcct
17521 gctgaaattg ttgacactgt gagtgctttg gtttatgata ataagcttaa agcacataaa
17581 gacaaatcag ctcaatgctt taaaatgttt tataagggtg ttatcacgca tgatgtttca
17641 tctgcaatta acaggccaca aataggcgtg gtaagagaat tccttacacg taaccctgct
17701 tggagaaaag ctgtctttat ttcaccttat aattcacaga atgctgtagc ctcaaagatt
17761 ttgggactac caactcaaac tgttgattca tcacagggct cagaatatga ctatgtcata
17821 ttcactcaaa ccactgaaac agctcactct tgtaatgtaa acagatttaa tgttgctatt
17881 accagagcaa aagtaggcat actttgcata atgtctgata gagaccttta tgacaagttg
17941 caatttacaa gtcttgaaat tccacgtagg aatgtggcaa ctttacaagc tgaaaatgta
18001 acaggactct ttaaagattg tagtaaggta atcactgggt tacatcctac acaggcacct
18061 acacacctca gtgttgacac taaattcaaa actgaaggtt tatgtgttga cgtacctggc
18121 atacctaagg acatgaccta tagaagactc atctctatga tgggttttaa aatgaattat
18181 caagttaatg gttaccctaa catgtttatc acccgcgaag aagctataag acatgtacgt
18241 gcatggattg gcttcgatgt cgaggggtgt catgctacta gagaagctgt tggtaccaat
18301 ttacctttac agctaggttt ttctacaggt gttaacctag ttgctgtacc tacaggttat
18361 gttgatacac ctaataatac agatttttcc agagttagtg ctaaaccacc gcctggagat

18421 caatttaaac acctcatacc acttatgtac aaaggacttc cttggaatgt agtgcgtata
18481 aagattgtac aaatgttaag tgacacactt aaaaatctct ctgacagagt cgtatttgtc
18541 ttatgggcac atggctttga gttgacatct atgaagtatt ttgtgaaaat aggacctgag
18601 cgcacctgtt gtctatgtga tagacgtgcc acatgctttt ccactgcttc agacacttat
18661 gcctgttggc atcattctat tggatttgat tacgtctata atccgtttat gattgatgtt
18721 caacaatggg gttttacagg taacctacaa agcaaccatg atctgtattg tcaagtccat
18781 ggtaatgcac atgtagctag ttgtgatgca atcatgacta ggtgtctagc tgtccacgag
18841 tgctttgtta agcgtgttga ctggactatt gaatatccta taattggtga tgaactgaag
18901 attaatgcgg cttgtagaaa ggttcaacac atggttgtta aagctgcatt attagcagac
18961 aaattcccag ttcttcacga cattggtaac cctaaagcta ttaagtgtgt acctcaagct
19021 gatgtagaat ggaagttcta tgatgcacag ccttgtagtg acaaagctta taaaatagaa
19081 gaattattct attcttatgc cacacattct gacaaattca cagatggtgt atgcctattt
19141 tggaattgca atgtcgatag atatcctgct aattccattg tttgtagatt tgacactaga
19201 gtgctatcta accttaactt gcctggttgt gatggtggca gtttgtatgt aaataaacat
19261 gcattccaca caccagcttt tgataaaagt gcttttgtta atttaaaaca attaccattt
19321 ttctattact ctgacagtcc atgtgagtct catggaaaac aagtagtgtc agacatagat
19381 tatgtaccac taaagtctgc tacgtgtata acacgttgca atttaggtgg tgctgtctgt
19441 agacatcatg ctaatgagta cagattgtat cttgatgctt ataacatgat gatctcagct
19501 ggctttagct tgtgggttta caaacaattt gatacttata acctctggaa cacttttaca
19561 agacttcaga gtttagaaaa tgtggctttt aatgttgtaa ataagggaca ctttgatgga
19621 caacagggtg aagtaccagt ttctatcatt aataacactg tttacacaaa agttgatggt
19681 gttgatgtag aattgtttga aaataaaaca acattacctg ttaatgtagc atttgagctt
19741 tgggctaagc gcaacattaa accagtacca gaggtgaaaa tactcaataa tttgggtgtg
19801 gacattgctg ctaatactgt gatctgggac tacaaaagag atgctccagc acatatatct
19861 actattggtg tttgttctat gactgacata gccaagaaac caattgaaac gatttgtgca
19921 ccactcactg tctttttga tggtagagtt gatggtcaag tagacttatt tagaaatgcc
19981 cgtaatggtg ttcttattac agagggtagt gttaaaggtt tacaaccatc tgtaggtccc
20041 aaacaagcta gtcttaatgg agtcacatta attggagaag ccgtaaaaac acagttcaat
20101 tattataaga aagttgatgg tgttgtccaa caattacctg aaacttactt tactcagagt
20161 agaaatttac aagaatttaa acccaggagt caaatggaaa ttgatttctt agaattagct
20221 atggatgaat tcattgaacg gtataaatta gaaggctatg ccttcgaaca tatcgtttat
20281 ggagatttta gtcatagtca gttaggtggt ttacatctac tgattggact agctaaacgt
20341 tttaaggaat cacctttga attagaagat tttattccta tggacagtac agttaaaaac
20401 tatttcataa cagatgcgca aacaggttca tctaagtgtg tgtgttctgt tattgattta
20461 ttacttgatg attttgttga aataataaaa tcccaagatt tatctgtagt ttctaaggtt
20521 gtcaaagtga ctattgacta tacagaaatt tcatttatgc tttggtgtaa agatggccat
20581 gtagaaacat tttacccaaa attacaatct agtcaagcgt ggcaaccggg tgttgctatg

20641 cctaatcttt acaaaatgca aagaatgcta ttagaaaagt gtgaccttca aaattatggt
20701 gatagtgcaa cattacctaa aggcataatg atgaatgtcg caaaatatac tcaactgtgt
20761 caatatttaa acacattaac attagctgta ccctataata tgagagttat acattttggt
20821 gctggttctg ataaaggagt tgcaccaggt acagctgttt taagacagtg gttgcctacg
20881 ggtacgctgc ttgtcgattc agatcttaat gactttgtct ctgatgcaga ttcaactttg
20941 attggtgatt gtgcaactgt acatacagct aataaatggg atctcattat tagtgatatg
21001 tacgacccta agactaaaaa tgttacaaaa gaaaatgact ctaaagaggg ttttttcact
21061 tacatttgtg ggtttataca acaaaagcta gctcttggag gttccgtggc tataaagata
21121 acagaacatt cttggagtgc tgatctttat aagctcatgg gacacttcgc atggtggaca
21181 gcctttgtta ctaatgtgaa tgcgtcatca tctgaagcat ttttaattgg atgtaattat
21241 cttggcaaac cacgcgaaca aatagatggt tatgtcatgc atgcaaatta catattttgg
21301 aggaatacaa atccaattca gttgtcttcc tattctttat ttgacatgag taaatttccc
21361 cttaaattaa ggggtactgc tgttatgtct ttaaaagaag gtcaaatcaa tgatatgatt
21421 ttatctcttc ttagtaaagg tagacttata attagagaaa acaacagagt tgttatttct
21481 agtgatgttc ttgttaacaa ctaaacgaac aatgtttgtt tttcttgttt tattgccact
21541 agtctctagt cagtgtgtta atcttataac cagaactcaa tcatacacta attctttcac
21601 acgtggtgtt tattaccctg acaaagtttt cagatcctca gttttacatt caactcagga
21661 cttgttctta ccttttcttt ccaatgttac ttggttccat gctatctctg ggaccaatgg
21721 tactaagagg tttgataacc ctgtcctacc atttaatgat ggtgtttatt ttgcttccac
21781 tgagaagtct aacataataa gaggctggat ttttggtact actttagatt cgaagaccca
21841 gtccctactt attgttaata acgctactaa tgttgttatt aaagtctgtg aatttcaatt
21901 ttgtaatgat ccatttttgg atgtttatta ccacaaaaac aacaaaagtt ggatggaaag
21961 tgagttcaga gtttattcta gtgcgaataa ttgcactttt gaatatgtct ctcagccttt
22021 tcttatggac cttgaaggaa aacagggtaa tttcaaaaat cttagggaat ttgtgtttaa
22081 gaatattgat ggttatttta aaatatattc taagcacacg cctattaatt tagggcgtga
22141 tctccctcag ggtttttcgg ctttagaacc attggtagat ttgccaatag gtattaacat
22201 cactaggttt caaactttac ttgctttaca tagaagttat ttgactcctg gtgattcttc
22261 ttcaggttgg acagctggtg ctgcagctta ttatgtgggt tatcttcaac ctaggacttt
22321 tctattaaaa tataatgaaa atggaaccat tacagatgct gtagactgtg cacttgaccc
22381 tctctcagaa acaaagtgta cgttgaaatc cttcactgta gaaaaaggaa tctatcaaac
22441 ttctaacttt agagtccaac caacagaatc tattgttaga tttcctaata ttacaaactt
22501 gtgccctttt gatgaagttt ttaacgccac cagatttgca tctgtttatg cttggaacag
22561 gaagagaatc agcaactgtg ttgctgatta ttctgtccta tataatttcg caccattttt
22621 cgcttttaag tgttatggag tgtctcctac taaattaaat gatctctgct ttactaatgt
22681 ctatgcagat tcatttgtaa ttagaggtaa tgaagtcagc caaatcgctc cagggcaaac
22741 tggaaatatt gctgattata attataaatt accagatgat tttacaggct gcgttatagc
22801 ttggaattct aacaagcttg attctaaggt tggtggtaat tataattacc ggtatagatt

22861 gtttaggaag tctaatctca aacctttga gagagatatt tcaactgaaa tctatcaggc
22921 cggtaacaaa ccttgtaatg gtgttgcagg tgttaattgt tactttcctt tacaatcata
22981 tggtttccga cccacttatg gtgttggtca ccaaccatac agagtagtag tactttcttt
23041 tgaacttcta catgcaccag caactgtttg tggacctaaa aagtctacta atttggttaa
23101 aaacaaatgt gtcaatttca acttcaatgg tttaacaggc acaggtgttc ttactgagtc
23161 taacaaaaag tttctgcctt tccaacaatt tggcagagac attgctgaca ctactgatgc
23221 tgtccgtgat ccacagacac ttgagattct tgacattaca ccatgttctt ttggtggtgt
23281 cagtgttata acaccaggaa caaatacttc taaccaggtt gctgttcttt atcagggtgt
23341 taactgcaca gaagtccctg ttgctattca tgcagatcaa cttactccta cttggcgtgt
23401 ttattctaca ggttctaatg tttttcaaac acgtgcaggc tgtttaatag gggctgaata
23461 tgtcaacaac tcatatgagt gtgacatacc cattggtgca ggtatatgcg ctagttatca
23521 gactcagact aagtctcatc ggcgggcacg tagtgtagct agtcaatcca tcattgccta
23581 cactatgtca cttggtgcag aaaattcagt tgcttactct aataactcta ttgccatacc
23641 cacaaatttt actattagtg ttaccacaga aattctacca gtgtctatga tcaagacatc
23701 agtagattgt acaatgtaca tttgtggtga ttcaactgaa tgcagcaatc ttttgttgca
23761 atatggcagt ttttgtacac aattaaaacg tgctttaact ggaatagctg ttgaacaaga
23821 caaaaacacc caagaagttt ttgcacaagt caaacaaatt tacaaaacac caccaattaa
23881 atattttggt ggttttaatt tttcacaaat attaccagat ccatcaaaac caagcaagag
23941 gtcatttatt gaagatctac ttttcaacaa agtgacactt gcagatgctg gcttcatcaa
24001 acaatatggt gattgccttg gtgatattgc tgctagagac ctcatttgtg cacaaaagtt
24061 taacggcctt actgttttgc caccttgct cacagatgaa atgattgctc aatacacttc
24121 tgcactgtta gcgggtacaa tcacttctgg ttggaccttt ggtgcaggtg ctgcattaca
24181 aataccattt gctatgcaaa tggcttatag gtttaatggt attggagtta cacagaatgt
24241 tctctatgag aaccaaaaat tgattgccaa ccaatttaat agtgctattg gcaaaattca
24301 agactcactt tcttccacag caagtgcact tggaaaactt caagatgtgg tcaaccataa
24361 tgcacaagct ttaaacacgc ttgttaaaca acttagctcc aaatttggtg caatttcaag
24421 tgttttaaat gatatccttt cacgtcttga caaagttgag gctgaagtgc aaattgatag
24481 gttgatcaca ggcagacttc aaagtttgca gacatatgtg actcaacaat taattagagc
24541 tgcagaaatc agatcttctg ctaatcttgc tgctactaaa atgtcagagt gtgtacttgg
24601 acaatcaaaa agagttgatt tttgtggaaa gggctatcat cttatgtcct tccctcagtc
24661 agcacctcat ggtgtagtct tcttgcatgt gacttatgtc cctgcacaag aaaagaactt
24721 cacaactgct cctgccattt gtcatgatgg aaaagcacac tttcctcgtg aaggtgtctt
24781 tgtttcaaat ggcacacact ggtttgtaac acaaaggaat ttttatgaac cacaaatcat
24841 tactacagac aacacatttg tgtctggtaa ctgtgatgtt gtaataggaa ttgtcaacaa
24901 cacagtttat gatcctttgc aacctgaatt agattcattc aaggaggagt tagataaata
24961 ttttaagaat catacatcac cagatgttga tttaggtgac atctctggca ttaatgcttc
25021 agttgtaaac attcaaaaag aaattgaccg cctcaatgag gttgccaaga atttaaatga

25081 atctctcatc gatctccaag aacttggaaa gtatgagcag tatataaaat ggccatggta
25141 catttggcta ggttttatag ctggcttgat tgccatagta atggtgacaa ttatgctttg
25201 ctgtatgacc agttgctgta gttgtctcaa gggctgttgt tcttgtggat cctgctgcaa
25261 atttgatgaa gacgactctg agccagtgct caaaggagtc aaattacatt acacataaac
25321 gaacttatgg atttgtttat gagaatcttc acaattggaa ctgtaacttt gaagcaaggt
25381 gaaatcaagg atgctactcc ttcagatttt gttcgcgcta ctgcaacgat accgatacaa
25441 gcctcactcc ctttcggatg gcttattgtt ggcgttgcac ttcttgctgt ttttcagagc
25501 gcttccaaaa tcataactct caaaaagaga tggcaactag cactctccaa gggtgttcac
25561 tttgtttgca acttgctgtt gttgtttgta acagtttact cacacctttt gctcgttgct
25621 gctggccttg aagccccttt tctctatctt tatgctttag tctacttctt gcagagtata
25681 aactttgtaa gaataataat gaggctttgg ctttgctgga aatgccgttc caaaaaccca
25741 ttactttatg atgccaacta ttttctttgc tggcatacta attgttacga ctattgtata
25801 ccttacaata gtgtaacttc ttcaattgtc attacttcag gtgatggcac aacaagtcct
25861 atttctgaac atgactacca gattggtggt tatactgaaa aatgggaatc tggagtaaaa
25921 gactgtgttg tattacacag ttacttcact tcagactatt accagctgta ctcaactcaa
25981 ttgagtacag acattggtgt tgaacatgtt accttcttca tctacaataa aattgttgat
26041 gagcctgaag aacatgtcca aattcacaca atcgacggtt catccggagt tgttaatcca
26101 gtaatggaac caatttatga tgaaccgacg acgactacta gcgtgccttt gtaagcacaa
26161 gctgatgagt acgaacttat gtactcattc gtttcggaag agataggtac gttaatagtt
26221 aatagcgtac ttctttttct tgctttcgtg gtattcttgc tagttacact agccatcctt
26281 actgcgcttc gattgtgtgc gtactgctgc aatattgtta acgtgagtct tgtaaaacct
26341 tctttttacg tttactctcg tgttaaaaat ctgaattctt ctagagttcc tgatcttctg
26401 gtctaaacga actaaatatt atattagttt ttctgtttgg aactttaatt ttagccatgg
26461 caaattccaa cggtactatt accgttgaag agcttaaaaa gctccttgaa gaatggaacc
26521 tagtaatagg tttcctattc cttacatgga tttgtcttct acaatttgcc tatgccaaca
26581 ggaataggtt tttgtatata attaagttaa ttttcctctg gctgttatgg ccagtaactt
26641 taacttgttt tgtgcttgct gctgtttaca gaataaattg gatcaccggt ggaattgcta
26701 tcgcaatggc ttgtcttgta ggcttgatgt ggctcagcta cttcattgct tctttcagac
26761 tgtttgcgcg tacgcgttcc atgtggtcat tcaatccaga aactaacatt cttctcaacg
26821 tgccactcca tggcactatt ctgaccagac cgcttctaga aagtgaactc gtaatcggag
26881 ctgtgatcct tcgtggacat cttcgtattg ctggacacca tctaggacgc tgtgacatca
26941 aggacctgcc taaagaaatc actgttgcta cgtcacgaac gctttcttat tacaaattgg
27001 gagcttcgca gcgtgtagca ggtgactcag gttttgctgc atacagtcgc tacaggattg
27061 gcaactataa attaaacaca gaccattcca gtagcagtga caatattgct ttgcttgtac
27121 agtaagtgac aacagatgtt tcatctcgtt gactttcagg ttactatagc agagatatta
27181 ctaattatta tgaggacttt taaagtttcc atttggaatc ttgattacat cataaacctc
27241 ataattaaaa atttatctaa gtcactaact gagaataaat attctcaatt agatgaagag

```
27301 caaccaatgg agattgatta aacgaacatg aaaattattc ttttcttggc actgataaca
27361 ctcgctactt gtgagcttta tcactaccaa gagtgtgtta gaggtacaac agtactttta
27421 aaagaacctt gctcttctgg aacatacgag ggcaattcac cattttatcc tctagctgat
27481 aacaaatttg cactgacttg ctttagcact caatttgctt ttgcttgtcc tgacggcgta
27541 aaacacgtct atcagttacg tgccagatca gtttcaccta aactgttcat cagacaagag
27601 gaagttcaag aactttactc tccaattttt cttattgttg cggcaatagt gtttataaca
27661 ctttgcttca cactcaaaag aaagacagaa tgattgaact ttcattaatt gacttctatt
27721 tgtgcttttt agcctttctg ttattccttg ttttaattat gcttattatc ttttggttct
27781 cacttgaact gcaagatcat aatgaaactt gtcacgccta aatgaacatg aaatttcttg
27841 ttttcttagg aannnnnnnn nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn nnacagtcat
27901 gtactcaaca tcaaccatat gtagttgatg acccgtgtcc tattcacttc tattctaaat
27961 ggtatattag agtaggagct agaaaatcag cacctttaat tgaattgtgc gtggatgagg
28021 ctggttctaa atcacccatt cagtacatcg atatcggtaa ttatacagtt tcctgtttac
28081 tttttacaat taattgccag gaacctaaat tgggtagtct tgtagtgcgt tgttcgttct
28141 atgaagactt tttagagtat catgacgttc gtgttgtttt agatttcatc taaacgaaca
28201 aacttaaatg tctgataatg gaccccaaaa tcagcgaaat gcactccgca ttacgtttgg
28261 tgggccctca gattcaactg gcagtaacca gaatggtggg gcgcgatcaa aacaacgtcg
28321 gccccagggt ttacccaata atactgcgtc ttggttcacc gctctcactc aacatggcaa
28381 ggaagacctt aaattccctc gaggacaagg cgttccaatt aacaccaata gcagtccaga
28441 tgaccaaatt ggctactacc gaagagctac cagacgaatt cgtggtggtg acggtaaaat
28501 gaaagatctc agtccaagat ggtatttcta ctacctagga actgggccag aagctggact
28561 tccctatggt gctaacaaag acggcatcat atgggttgca actgagggag ccttgaatac
28621 accaaaagat cacattggca cccgcaatcc tgctaacaat gctgcaatcg tgctacaact
28681 tcctcaagga acaacattgc caaaaggctt ctacgcagaa gggagcagag gcggcagtca
28741 agcctcttct cgttcctcat cacgtagtcg caacagttca agaaattcaa ctccaggcag
28801 cagtaaacga acttctcctg ctagaatggc tggcaatggc ggtgatgctg ctcttgcttt
28861 gctgctgctt gacagattga accagcttga gagcaaaatg tctggtaaag gccaacaaca
28921 acaaggccaa actgtcacta agaaatctgc tgctgaggct tctaagaagc ctcggcaaaa
28981 acgtactgcc actaaagcat acaatgtaac acaagctttc ggcagacgtg gtccagaaca
29041 aacccaagga aattttgggg accaggaact aatcagacaa ggaactgatt acaaacattg
29101 gccgcaaatt gcacaatttg cccccagcgc ttcagcgttc ttcggaatgt cgcgcattgg
29161 catggaagtc acaccttcgg gaacgtggtt gacctacaca ggtgccatca aattggatga
29221 caaagatcca aatttcaaag atcaagtcat tttgctgaat aagcatattg acgcatacaa
29281 aacattccca ccaacagagc ctaaaaagga caaaaagaag aaggctgatg aaactcaagc
29341 cttaccgcag agacagaaga aacagcaaac tgtgactctt cttcctgctg cagatttgga
29401 tgatttctcc aaacaattgc aacaatccat gagccgtgct gactcaactc aggcctaaac
29461 tcatgcagnn nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn nncgcttttc cgtttacgat
```

```
    29521 atatagtcta ctcttgtgca gaatgaattc tcgtaactac atagcacaag tagatgtagt
    29581 taactttaat ctcacatagc aatctttaat cagtgtgtaa cattagggag gacttgaaag
    29641 agccaccaca ttttcaccta cagtgaacaa tgctagggag agctg
//
```

Similarly, apart from the term, we can use the chaining method using regular expression. This lists the IDs associated with accD (the gene name) and the organism E. Coli:

In [42]:
```
handle = Entrez.esearch(db='nucleotide', term='accD[Gene Name] AND "E. coli"[Organism]', retmax="20")
result_list = Entrez.read(handle)
```

In [43]:
```
id_list = result_list['IdList']
count = result_list['Count']

print(id_list)
print("\n")
print(count)
```

```
['2540286096', '2540285939', '2540285880', '2540285612', '2540285515', '2540285302', '2540
285285', '2540281271', '2540281265', '2540281264', '2536864279', '2535150858', '253515085
7', '2535150855', '2535150854', '2535150853', '2535150852', '2535150851', '2535150850', '2
535150844']


220644
```

In [44]:
```
handle.close()
```