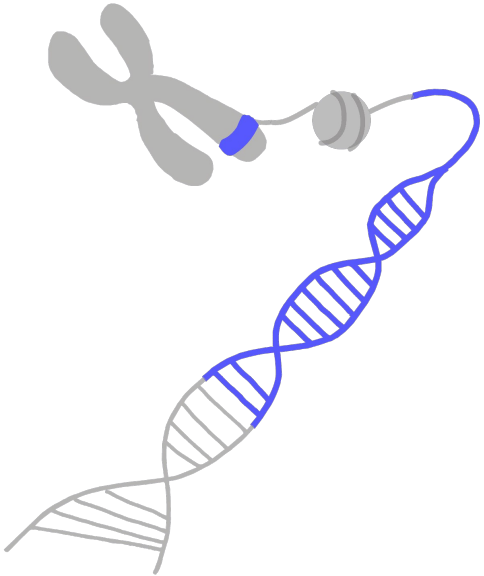
Análisis de la información de los archivos FASTA y GenBank

- Gen CFTR (Cystic Fibrosis Transmembrane Conductance)
- Proteína 4INS (Insulina Humana)



Gen CFTR

- Fundamental en el estudio de la fibrosis quística
- Gen específico

Datos en formato FASTA


```
fasta_seq = SeqIO.read(fasta_file_path, "fasta")
```

```
fasta_info = {
    "ID": fasta_seq.id,
    "Description": fasta_seq.description,
    "Sequence Length": len(fasta_seq.seq),
    "Composition": {
        "A": fasta_seq.seq.count("A"),
        "T": fasta_seq.seq.count("T"),
        "C": fasta_seq.seq.count("C"),
        "G": fasta_seq.seq.count("G")
    }
}
```

- **ID:** 'NM_000492.3',
- **Description:** 'NM_000492.3 Homo sapiens CF transmembrane conductance regulator (CFTR) mRNA'
- **Sequence Length:** 6132
- **Composition:**
  - 'A': 1887
  - 'T': 1731
  - 'C': 1182
  - 'G': 1332

Datos en formato GenBank

```python
genbank_seq = SeqIO.read(genbank_file_path, "genbank")
```

```python
genbank_info = {
    "ID": genbank_seq.id,
    "Description": genbank_seq.description,
    "Sequence Length": len(genbank_seq.seq),
    "Composition": {
        "A": genbank_seq.seq.count("A"),
        "T": genbank_seq.seq.count("T"),
        "C": genbank_seq.seq.count("C"),
        "G": genbank_seq.seq.count("G")
    },
    "Annotations": genbank_seq.annotations,
    "Features": genbank_seq.features,
    "Origin": genbank_seq.seq

}
```

- **Referencias bibliográficas:** referencias científicas con el gen CFTR, de gran relevancia, ej: Reference(title='Domain-interface dynamics of CFTR revealed by stabilizing nanobodies', ...), Reference(title='Unusual Cystic Fibrosis Transmembrane Conductance Regulator Mutations and Liver Disease: A Case Series and Review of the Literature', ...)
- **Comentarios:** basado en RNA-seq, criterios de selección RefSeq.

- **Características generales:** elementos importantes de la secuencia:
  - región del gen
  - PRI
  - posiciones exones
  - 27 agosto 2019
  - regiones de interés
  - sitios regulatorios
  - estructura del gen
- **Accesiones y versión:** Versión 3 de la secuencia que ha sido reemplazada por versión 4
- **Origin:** secuencia de nucleótidos completa del gen
- **Palabras clave:** RefSeq y RefSeq Select (revisada por NCBI)
- **Resumen funcional:** regulación del transporte de iones, relación con mutaciones que causan fibrosis quística
- **Fuente y taxonomía:** Homo sapiens, Vertebrata, Mammalia, Hominidae, Metazoa, Chordata, Craniata, etc.

Proteína 4INS

- Estructura de la insulina de cerdo
- Formada por 2 cadenas A y B
- Cadenas específicas de la proteína de la insulina

Datos en formato FASTA


```
fasta_seq = SeqIO.read(fasta_file_path, "fasta")
```

```
for seq_record in fasta_sequences:
    fasta_info = {
        "ID": seq_record.id,
        "Description": seq_record.description,
        "Sequence Length": len(seq_record.seq),
        "Composition": {
            "A": seq_record.seq.count("A"),
            "C": seq_record.seq.count("C"),
            "D": seq_record.seq.count("D"),
            "E": seq_record.seq.count("E"),
            "F": seq_record.seq.count("F"),
            "G": seq_record.seq.count("G"),
            "H": seq_record.seq.count("H"),
            "I": seq_record.seq.count("I"),
            "K": seq_record.seq.count("K"),
            "L": seq_record.seq.count("L"),
            "M": seq_record.seq.count("M"),
            "N": seq_record.seq.count("N"),
            "P": seq_record.seq.count("P"),
            "Q": seq_record.seq.count("Q"),
            "R": seq_record.seq.count("R"),
            "S": seq_record.seq.count("S"),
            "T": seq_record.seq.count("T"),
            "V": seq_record.seq.count("V"),
            "W": seq_record.seq.count("W"),
            "Y": seq_record.seq.count("Y")
        }
    }
```

- **ID:** pdb|4INS|A,
- **Description:** pdb|4INS|A Chain A, INSULIN (CHAIN A)
- **Sequence Length:** 21
- **Composition:** 'A': 0, 'C': 2, 'D': 0, 'E': 2, 'F': 0, 'G': 3, 'H': 0, 'I': 0, 'K': 0, 'L': 2, 'M': 0, 'N': 2, 'P': 0, 'Q': 2, 'R': 0, 'S': 2, 'T': 1, 'V': 3, 'W': 0, 'Y': 2

Datos en formato GenPept

```
genbank_seq = SeqIO.read(genbank_file_path, "genbank")
```

```
for seq_record in genpept_sequences:
    genpept_info = {
        "ID": seq_record.id,
        "Description": seq_record.description,
        "Sequence Length": len(seq_record.seq),
        "Amino Acid Composition": {
            "A": seq_record.seq.count("A"),
            "C": seq_record.seq.count("C"),
            "D": seq_record.seq.count("D"),
            "E": seq_record.seq.count("E"),
            "F": seq_record.seq.count("F"),
            "G": seq_record.seq.count("G"),
            "H": seq_record.seq.count("H"),
            "I": seq_record.seq.count("I"),
            "K": seq_record.seq.count("K"),
            "L": seq_record.seq.count("L"),
            "M": seq_record.seq.count("M"),
            "N": seq_record.seq.count("N"),
            "P": seq_record.seq.count("P"),
            "Q": seq_record.seq.count("Q"),
            "R": seq_record.seq.count("R"),
            "S": seq_record.seq.count("S"),
            "T": seq_record.seq.count("T"),
            "V": seq_record.seq.count("V"),
            "W": seq_record.seq.count("W"),
            "Y": seq_record.seq.count("Y"),
        },
        "Annotations": seq_record.annotations,
        "Features": seq_record.features,
        "Origin": seq_record.seq
    }
```

- **Anotaciones y comentario:**
  - **Palabras clave:** Insulina (estructural)
  - **Comentario:** Estructura cristalina de insulina 2ZN a 1.5 Å de resolución
- **Referencias destacadas:**
  - **Fecha:** 25 octubre 2024
  - **Estructura de la insulina 2ZN:** análisis cristalográfico
  - **Métodos de refinamiento:** evaluación comparativa en modelos atómicos
- **Accesiones y fuente:**
  - **Accesiones:** 4INS_A
  - **Relaciones estructurales:** hexámero de insulina coordinado con zinc
  - **Fuente de datos:** pdb: molecula 4INS, cadena A
- **Features:**
  - **Método experimental:** Difracción de rayos X
  - **Ubicación fuente:** posición 0-21
  - **Organismo:** Sus scrofa (cerdo)
  - **Estructura secundaria:** posiciones 0-9
  - **Enlaces:** posiciones 5-6 y 10-11
- **Origen:**
  - **Taxonomía:** Eukaryota, Metazoa, Chordata, Craniata, Vertebrata, Mammalia, Laurasiatheria, Suina, Sus.
  - **Secuencia proteica:** GIVEQCCTSICSLYQLENYCN

INTERPRETACIÓN 2º ENLACE

NIH National Library of Medicine
National Center for Biotechnology Information

Nucleotide [Nucleotide] [ ]
Advanced

GenBank

Arabidopsis thaliana DNA chromosome 4, c

GenBank: AL161505.2
FASTA Graphics

Go to:

```
LOCUS       AL161505              198176 bp    DNA     linear
DEFINITION  Arabidopsis thaliana DNA chromosome 4, contig fragme
ACCESSION   AL161505
VERSION     AL161505.2
KEYWORDS    .
SOURCE      Arabidopsis thaliana (thale cress)
  ORGANISM  Arabidopsis thaliana
            Eukaryota; Viridiplantae; Streptophyta; Embryophyta;
            Spermatophyta; Magnoliopsida; eudicotyledons; Gunner
            Pentapetalae; rosids; malvids; Brassicales; Brassica
            Camelineae; Arabidopsis.
REFERENCE   1  (bases 1 to 1; 63093 to 175721)
  AUTHORS   Wilson,R., Lamar,B., Stoneking,T., Stumpf,J., Mewes,
            and Mayer,K.F.X.
  JOURNAL   Unpublished
REFERENCE   2  (bases 1 to 198176)
  AUTHORS   EU Arabidopsis sequencing,project.
  TITLE     Direct Submission
  JOURNAL   Submitted (10-MAR-2000) MIPS, at the Max-Planck-Inst
            Biochemie, Am Klopferspitz 18a, D-82152 Martinsried,
            lemcke@mips.biochem.mpg.de,mayer@mips.biochem.mpg.de
            Coordinator: Mike Bevan, Molecular Genetics Departme
            Laboratory, John Innes Centre, Colney Lane, NR4 7UJ
            E-mail: michael.bevan@bbsrc.ac.uk
COMMENT     Information on performance of analysis and a more de
            annotation of this entry and other sequences of chro
            and 5 can be viewed at: http://www.mips.biochem.mpg.
            this fragment has an overlap with ATCHRIV16 at the 5
            overlap with ATCHRIV18 at the 3' end.
FEATURES             Location/Qualifiers
     source          1..198176
                     /organism="Arabidopsis thaliana"
                     /mol_type="genomic DNA"
                     /db_xref="taxon:3702"
                     /chromosome="4"
                     /ecotype="Columbia"
```

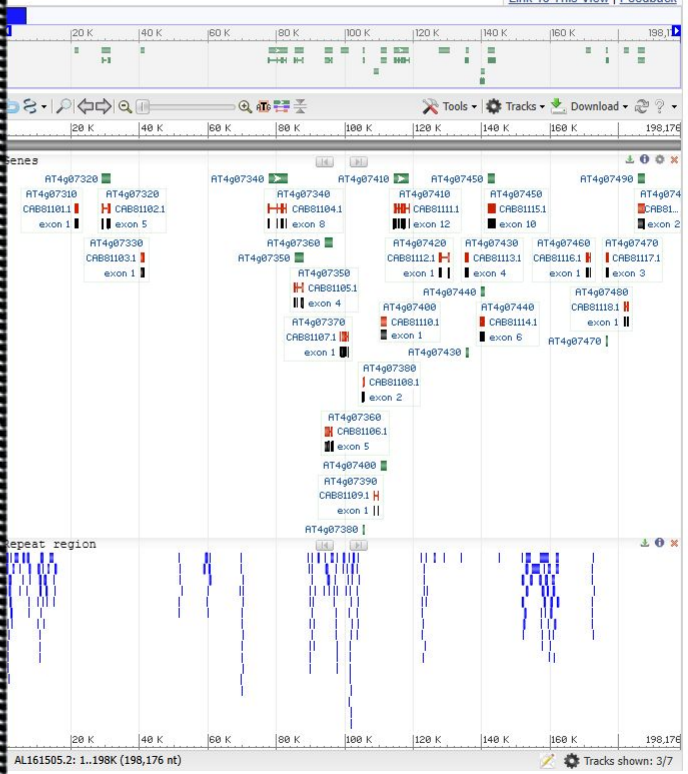
```
misc_feature    12310..15735
                /note="pseudogene, similar to retrovirus-related
                polyproteins"
repeat_region   12735..13052
                /note="T32N15_del_retroTn"
repeat_region   13053..13086
                /note="T32N15_del_retroTn"
repeat_region   13087..13862
                /note="T32N15_del_retroTn"
repeat_region   13864..14927
                /note="T32N15_del_retroTn"
repeat_region   14920..15027
                /note="T32N15_del_retroTn"
repeat_region   15024..15226
                /note="T32N15_del_retroTn"
repeat_region   15224..15350
                /note="T32N15_del_retroTn"
repeat_region   15340..15739
                /note="T32N15_del_retroTn"
gene            21183..22195
                /gene="AT4g07310"
CDS             join(21183..21320,21398..21662,21795..21897,21973..22195)
                /gene="AT4g07310"
                /note="contains similarity to Arabidopsis thaliana
                hypothetical proteins, see GB:AF058826;
                similarity to"
                /codon_start=1
                /protein_id="CAB81101.1"
                /db_xref="InterPro:IPR003871"
                /db_xref="InterPro:IPR012340"
                /db_xref="InterPro:IPR016027"
                /db_xref="UniProtKB/TrEMBL:Q9S9Y1"
                /translation="MESFHPLSLLKPSIRGWCIRGRVVRTFLVSLVPSSKVMGLILAE
                EHGMTIEATVGYKMSDHYKDFINEGEWTITNFGVVENSGSVRATTHSFKIGFFVDIV
                VRLTSPLPAIPHYRLASFSSIIDDEIDKSVLVDLVGAIYDVGELINTRPKQNNVDDLT
                LIFKISDNENRVLECLATKKEALDFDHNYRRYGGGVIVAVLGWWKIDRYFDGPKNVRV
                CTAGTISTVFPDPDPIPESNEIHEI"
exon            21183..21320
                /gene="AT4g07310"
                /number=1
intron          21321..21397
                /gene="AT4g07310"
                /number=1
exon            21398..21662
                /gene="AT4g07310"
                /number=2
intron          21663..21794
                /gene="AT4g07310"
                /number=2
exon            21795..21897
                /gene="AT4g07310"
                /number=3
intron          21898..21972
                /gene="AT4g07310"
                /number=3
```

Graphics Send to:

Arabidopsis thaliana DNA chromosome 4, contig fragment No. 17

GenBank: AL161505.2
GenBank FASTA

Link To This View | Feedback



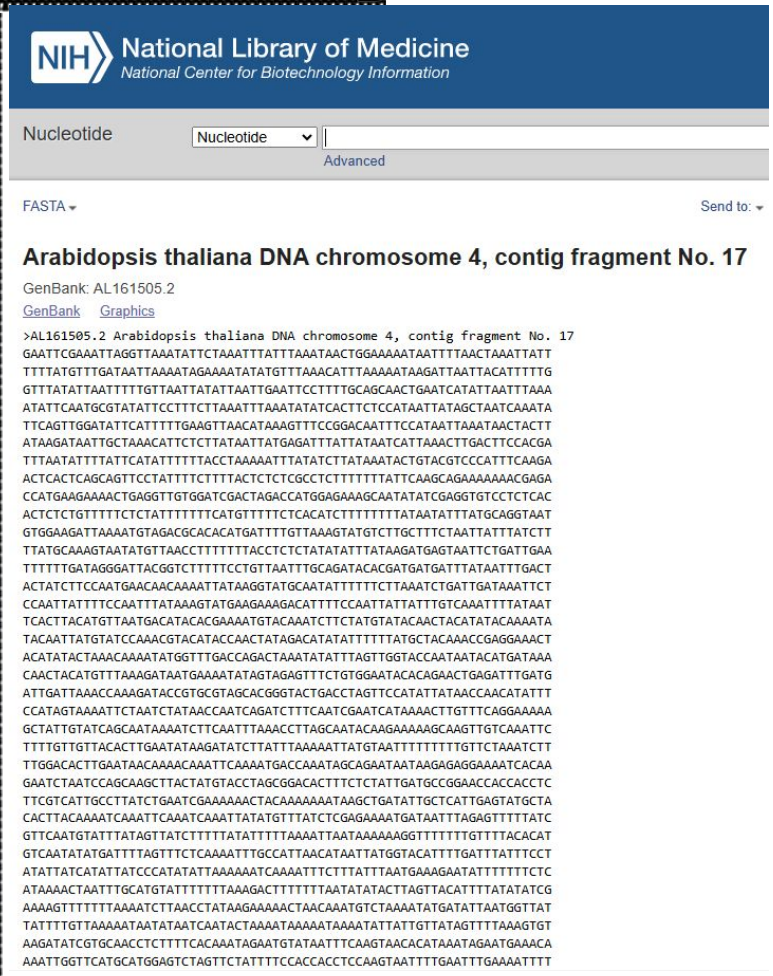
Analyze this sequence
Run BLAST
Pick Primers

Related information
Protein
Taxonomy
Component Of
Full text in PMC
Gene
PubMed (Weighted)
GEO Profiles

Recent activity
Turn Off Clear
Arabidopsis thaliana DNA chromosome 4, contig fra(Nucleotide
1A3N (10) Nucleotide
4INS (5) Protein
Chain A, GREEN FLUORESCENT PROTEIN Protein
1EMA (3) Protein
See more...

INTERPRETACIÓN 1º ENLACE

```
     intron          complement(186935..187163)
                     /number=1
     exon            complement(187164..187579)
                     /gene="AT4g07490"
                     /number=2
     misc_feature    188252..191155
                     /note="contains similarity to athila retroelement
                     ORF1-like proteins"
     misc_feature    192885..195541
                     /note="pseudogene, contains similarity to athila
                     retroelement ORF1-like proteins"
     misc_feature    196274..198176
                     /note="pseudogene"
ORIGIN
        1 gaattcgaaa ttaggttaaa tattctaaat ttatttaaat aactggaaaa ataattttaa
       61 ctaaattatt ttttatgttt gataatttaaa atagaaaata tatgtttaaa catttaaaaa
      121 taagattaat tacatttttg gtttatatta atttttgtta attatattaa ttgaattcct
      181 tttgcagcaa ctgaatcata ttaatttaaa atattcaatg cgtatattcc tttcttaaat
      241 ttaaatatat cacttctcca taattatagc taatcaaata ttcagttgga tattcatttt
      301 tgaagttaac ataaagtttc cggacaattt ccataattaa ataactactt ataagataat
      361 tgctaaacat tctcttataa ttatgagatt tattataatc attaaacttg acttccacga
      421 tttaatattt tattcatatt ttttacctaa aaatttatat cttatataaa ctgtacgtcc
      481 catttcaaga actcactcag cagttcctat tttcttttac tctctcgcct cttttttttt
      541 tcaagcagaa aaaaacgaga ccatgaagaa aactgaggtt gtggatcgac tagaccatgg
      601 agaaagcaat atatcgaggt gtcctctcac actctctgtt tttctctatt tttttcatgt
      661 ttttctcaca tctttttttt ataatattta tgcaggtaat gtggaagatt aaaatgtaga
      721 cgcacacatg attttgttaa agtatgtctt gctttctaat tatttatctt ttatgcaaag
      781 taatatgtta acctttttttt acctctctat atatttataa gatgagtaat tctgattgaa
      841 ttttttgata gggattacgg tctttttcct gttaatttgc agatacacga tgatgattta
      901 taatttgact actatcttcc aatgaacaac aaaattataa ggtatgcaat attttttctt
      961 aaatctgatt gataaattct ccaatttattt tccaatttat aaagtatgaa gaaagacatt
     1021 ttccaattat tatttgtcaa attttataat tcacttacat gttaatgaca tacacgaaaa
     1081 tgtacaaatc ttctatgtat acaactacat atacaaaata tacaattatg tatccaaacg
     1141 tacataccaa ctatagacat atatttttta tgctacaaac cgaggaaact acatatacta
     1201 aacaaaatat ggtttgacca gactaaatat atttagttgg taccaataat acatgataaa
     1261 caactacatg tttaaagata atgaaaatat agtagagttt ctgtggaata cacagaactg
     1321 agatttgatg attgattaaa ccaaagatac cgtgcgtagc acgggtactg acctagttcc
     1381 atattataac caacatattt ccatagtaaa attctaatct ataaccaatc agatctttca
     1441 atcgaatcat aaaacttgtt tcaggaaaaa gctattgtat cagcaataaa atcttcaatt
     1501 taaaccttag caatacaaga aaaagcaagt tgtcaaattc ttttgttgtt acacttgaat
     1561 ataagatatc ttatttaaaa attatgtaat ttttttttgt tctaaatctt ttggacactt
     1621 gaataacaaa acaaattcaa aatgaccaaa tagcagaata ataagagagg aaaatcacaa
     1681 gaatctaatc cagcaagctt actatgtacc tagcggacac tttctctatt gatgccggaa
     1741 ccaccacctc ttcgtcattg ccttatctga atcgaaaaaa ctacaaaaaa ataagctgat
     1801 attgctcatt gagtatgcta cacttacaaa atcaaattca aatcaaatta tatgtttatc
     1861 tcgagaaaat gataatttag agtttttatc gttcaatgta tttatagtta tctttttata
     1921 tttttaaaat taataaaaaa ggtttttttg ttttacacat gtcaatatat gattttagtt
     1981 tctcaaaatt tgccattaac ataattatgg tacattttga tttatttcct atattatcat
     2041 attatcccat atattaaaaa atcaaaattt ctttatttaa tgaaagaata ttttttctc
     2101 ataaaactaa tttgcatgta ttttttttaa gacttttttt aatatatact tagttacatt
     2161 ttatatatcg aaaagttttt ttaaaatctt aacctataag aaaaactaac aaatgtctaa
     2221 aatatgatat taatggttat tattttgtta aaaataatat aatcaatact aaaataaaaa
     2281 taaaatatta ttgttatagt tttaaagtgt aagatatcgt gcaacctctt ttcacaaata
     2341 gaatgtataa tttcaagtaa cacataaata gaatgaaaca aaattggttc atgcatggag
     2401 tctagttcta ttttccacca cctccaagta attttgaatt tgaaaatttt gtttttattt
```



National Library of Medicine
National Center for Biotechnology Information

Nucleotide

Nucleotide

Advanced

FASTA

Send to:

Arabidopsis thaliana DNA chromosome 4, contig fragment No. 17

GenBank: AL161505.2

GenBank Graphics

```
>AL161505.2 Arabidopsis thaliana DNA chromosome 4, contig fragment No. 17
GAATTCGAAATTAGGTTAAATATTCTAAATTTATTTAAATAACTGGAAAAATAATTTTAACTAAATTATT
TTTTATGTTTGATAATTAAAATAGAAAAATATATGTTTAAACATTTAAAAATAAGATTAATTACATTTTTG
GTTTATATTAATTTTTGTTAATTATATTAATTGAATTCCTTTTGCAGCAACTGAATCATATTAATTTAAA
ATATTCAATGCGTATATTCCTTTCTTAAATTTAAATATATCACTTCTCCATAATTATAGCTAATCAAATA
TTCAGTTGGATATTCATTTTTGAAGTTAACATAAAGTTTCCGGACAATTTCCATAATTAAATAACTACTT
ATAAGATAATTGCTAAACATTCTCTTATAATTATGAGATTTATTATAATCATTAAACTTGACTTCCACGA
TTTAATATTTTATTCATATTTTTTACCTAAAAATTTATATCTTATAAATACTGTACGTCCCATTTCAAGA
ACTCACTCAGCAGTTCCTATTTTCTTTTACTCTCTCGCCTCTTTTTTTATTCAAGCAGAAAAAAACGAGA
CCATGAAGAAAACTGAGGTTGTGGATCGACTAGACCATGGAGAAAGCAATATATCGAGGTGTCCTCTCAC
ACTCTCTGTTTTTCTCTATTTTTTTTCATGTTTTTCTCACATCTTTTTTTTATAATATTTATGCAGGTAAT
GTGGAAGATTAAAATGTAGACGCACACATGATTTTGTTAAAGTATGTCTTGCTTTCTAATTATTTATCTT
TTATGCAAAGTAATATGTTAACCTTTTTTTACCTCTCTATATATTTATAAGATGAGTAATTCTGATTGAA
TTTTTTGATAGGGATTACGGTCTTTTTCCTGTTAATTTGCAGATACACGATGATGATTTATAATTTGACT
ACTATCTTCCAATGAACAACAAAATTATAAGGTATGCAATATTTTTTCTTAAATCTGATTGATAAATTCT
CCAATTATTTTCCAATTTATAAAGTATGAAGAAAGACATTTTCCAATTATTATTTGTCAAATTTTATAAT
TCACTTACATGTTAATGACATACACGAAAATGTACAAATCTTCTATGTATACAACTACATATACAAAATA
TACAATTATGTATCCAAACGTACATACCAACTATAGACATATATTTTTTATGCTACAAACCGAGGAAACT
ACATATACTAAACAAAATATGGTTTGACCAGACTAAATATATTTAGTTGGTACCAATAATACATGATAAA
CAACTACATGTTTAAAGATAATGAAAATATAGTAGAGTTTCTGTGGAATACACAGAACTGAGATTTGATG
ATTGATTAAACCAAAGATACCGTGCGTAGCACGGGTACTGACCTAGTTCCATATTATAACCAACATATTT
CCATAGTAAAATTCTAATCTATAACCAATCAGATCTTTCAATCGAATCATAAAACTTGTTTCAGGAAAAA
GCTATTGTATCAGCAATAAAATCTTCAATTTAAACCTTAGCAATACAAGAAAAAGCAAGTTGTCAAATTC
TTTTGTTGTTACACTTGAATATAAGATATCTTATTTAAAAATTATGTAATTTTTTTTTGTTCTAAATCTT
TTGGACACTTGAATAACAAAACAAATTCAAAATGACCAAATAGCAGAATAATAAGAGAGGAAAATCACAA
GAATCTAATCCAGCAAGCTTACTATGTACCTAGCGGACACTTTCTCTATTGATGCCGGAACCACCACCTC
TTCGTCATTGCCTTATCTGAATCGAAAAAACTACAAAAAAATAAGCTGATATTGCTCATTGAGTATGCTA
CACTTACAAAATCAAATTCAAATCAAATTATATGTTTATCTCGAGAAAATGATAATTTAGAGTTTTTATC
GTTCAATGTATTTATAGTTATCTTTTTATATTTTTAAAATTAATAAAAAAGGTTTTTTTGTTTTACACAT
GTCAATATATGATTTTAGTTTCTCAAAATTTGCCATTAACATAATTATGGTACATTTTGATTTATTTCCT
ATATTATCATATTATCCCATATATTAAAAAATCAAAATTTCTTTATTTAATGAAAGAATATTTTTTTCTC
ATAAAACTAATTTGCATGTATTTTTTTAAAGACTTTTTTTAATATATACTTAGTTACATTTTATATATCG
AAAAGTTTTTTTAAAATCTTAACCTATAAGAAAAACTAACAAATGTCTAAAATATGATATTAATGGTTAT
TATTTTGTTAAAAATAATATAATCAATACTAAAATAAAAATAAAATATTATTGTTATAGTTTTAAAGTGT
AAGATATCGTGCAACCTCTTTTCACAAATAGAATGTATAATTTCAAGTAACACATAAATAGAATGAAACA
AAATTGGTTCATGCATGGAGTCTAGTTCTATTTTCCACCACCTCCAAGTAATTTTGAATTTGAAAATTTT
```