Expresiones regulares

Bioinformática
Facultad de Ciencias e Ingeniería

Expresión regular (regex)

- Cadena de caracteres que define un patrón que se encuentra en un bloque de texto.

- Se puede construir una expresión regular para permitir múltiples coincidencias posibles, mientras se restringen otras posibilidades.

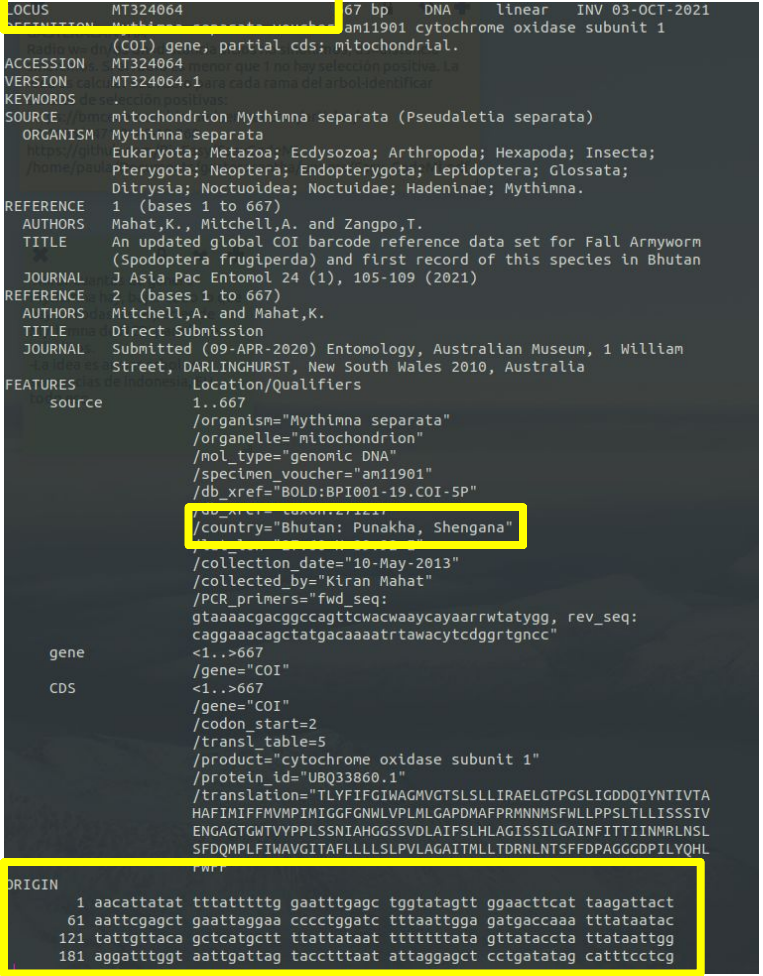
A la clase de profundización en bioinformática del periodo 2025-1 asisten 10 estudiantes.

A la clase de profundización en Bioinformática del periodo 2025-1 asisten 10 estudiantes.

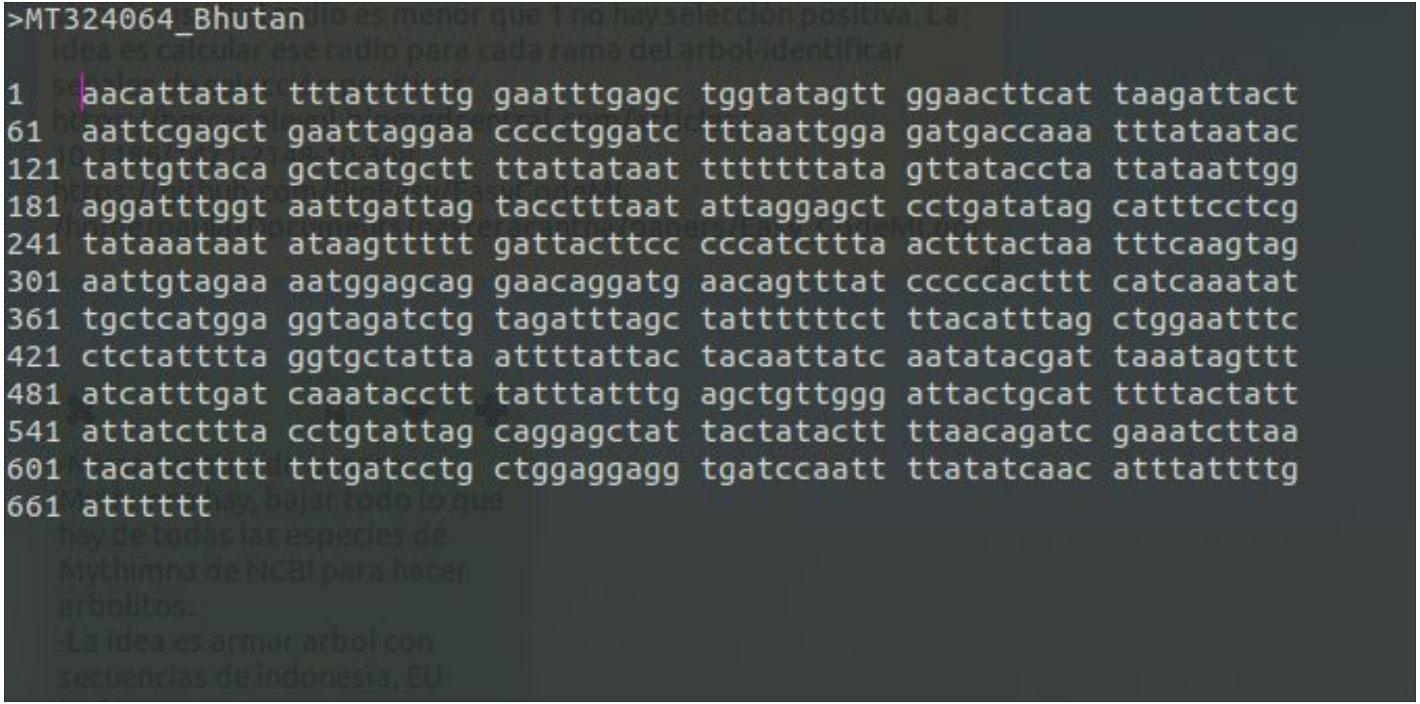
A Bioinformática vienen 10 estudiantes.

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) reference genome

SARS-CoV-2 genome

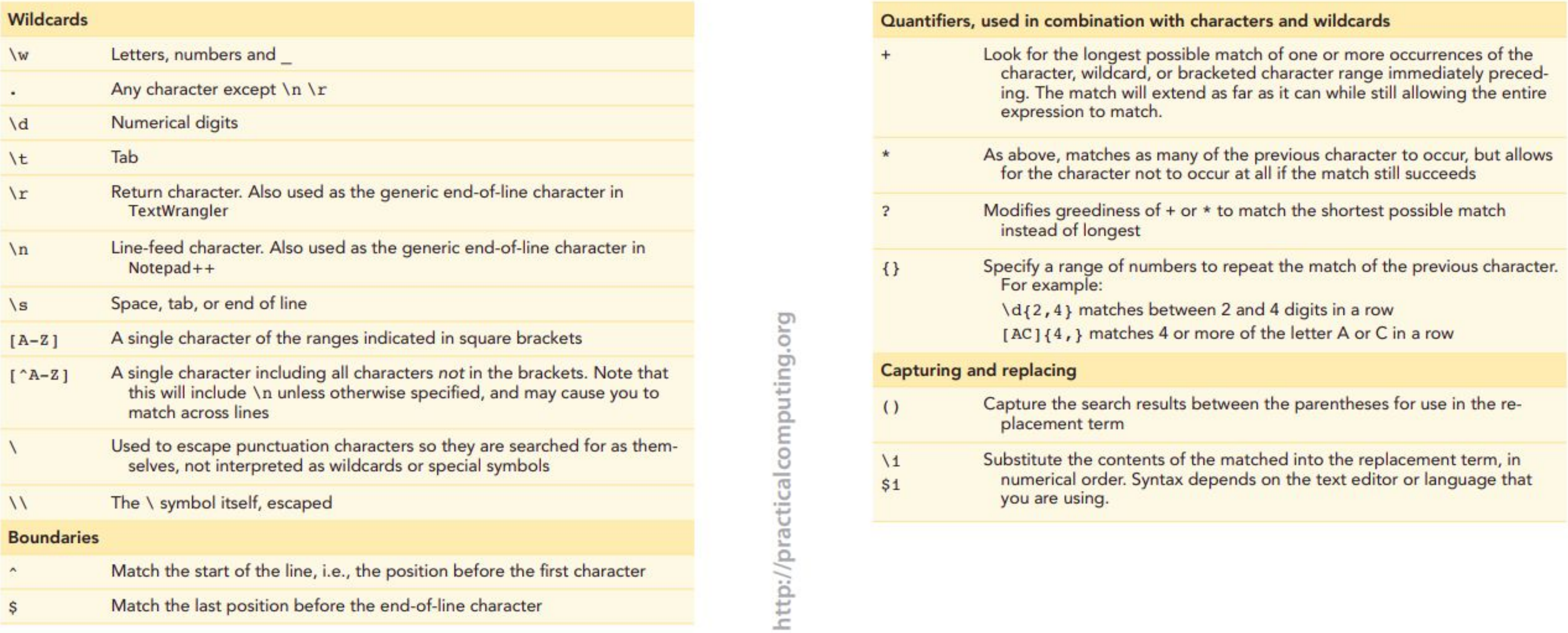


```
LOCUS       MT324064                 667 bp    DNA     linear   INV 03-OCT-2021
DEFINITION  Mythimna separata voucher am11901 cytochrome oxidase subunit 1
            (COI) gene, partial cds; mitochondrial.
ACCESSION   MT324064
VERSION     MT324064.1
KEYWORDS    .
SOURCE      mitochondrion Mythimna separata (Pseudaletia separata)
  ORGANISM  Mythimna separata
            Eukaryota; Metazoa; Ecdysozoa; Arthropoda; Hexapoda; Insecta;
            Pterygota; Neoptera; Endopterygota; Lepidoptera; Glossata;
            Ditrysia; Noctuoidea; Noctuidae; Hadeninae; Mythimna.
REFERENCE   1  (bases 1 to 667)
  AUTHORS   Mahat,K., Mitchell,A. and Zangpo,T.
  TITLE     An updated global COI barcode reference data set for Fall Armyworm
            (Spodoptera frugiperda) and first record of this species in Bhutan
  JOURNAL   J Asia Pac Entomol 24 (1), 105-109 (2021)
REFERENCE   2  (bases 1 to 667)
  AUTHORS   Mitchell,A. and Mahat,K.
  TITLE     Direct Submission
  JOURNAL   Submitted (09-APR-2020) Entomology, Australian Museum, 1 William
            Street, DARLINGHURST, New South Wales 2010, Australia
FEATURES             Location/Qualifiers
     source          1..667
                     /organism="Mythimna separata"
                     /organelle="mitochondrion"
                     /mol_type="genomic DNA"
                     /specimen_voucher="am11901"
                     /db_xref="BOLD:BPI001-19.COI-5P"
                     /db_xref="taxon:27121"
                     /country="Bhutan: Punakha, Shengana"
                     /lat_lon="27.63 N 89.82 E"
                     /collection_date="10-May-2013"
                     /collected_by="Kiran Mahat"
                     /PCR_primers="fwd_seq:
                     gtaaaacgacggccagttcwacwaaycayaarrwtatygg, rev_seq:
                     caggaaacagctatgacaaaatrtawacytcdggrtgncc"
     gene            <1..>667
                     /gene="COI"
     CDS             <1..>667
                     /gene="COI"
                     /codon_start=2
                     /transl_table=5
                     /product="cytochrome oxidase subunit 1"
                     /protein_id="UBQ33860.1"
                     /translation="TLYFIFGIWAGMVGTSLSLLIRAELGTPGSLIGDDQIYNTIVTA
                     HAFIMIFFMVMPIMIGGFGNWLVPLMLGAPDMAFPRMNNMSFWLLPPSLTLLISSSIV
                     ENGAGTGWTVYPPLSSNIAHGGSSVDLAIFSLHLAGISSILGAINFITTIINMRLNSL
                     SFDQMPLFIWAVGITAFLLLLSLPVLAGAITMLLTDRNLNTSFFDPAGGGDPILYQHL
                     FWFF
ORIGIN
        1 aacattatat tttatttttg gaatttgagc tggtatagtt ggaacttcat taagattact
       61 aattcgagct gaattaggaa cccctggatc tttaattgga gatgaccaaa tttataatac
      121 tattgttaca gctcatgctt ttattataat ttttttttata gttatacctta ttataattgg
      181 aggatttggt aattgattag tacctttaat attaggagct cctgatatag catttcctcg
```



```
>MT324064_Bhutan

1   aacattatat tttatttttg gaatttgagc tggtatagtt ggaacttcat taagattact
61  aattcgagct gaattaggaa cccctggatc tttaattgga gatgaccaaa tttataatac
121 tattgttaca gctcatgctt ttattataat tttttttata gttataccta ttataattgg
181 aggatttggt aattgattag tacctttaat attaggagct cctgatatag catttcctcg
241 tataaataat ataagttttt gattacttcc cccatcttta actttactaa tttcaagtag
301 aattgtagaa aatggagcag gaacaggatg aacagtttat cccccacttt catcaaatat
361 tgctcatgga ggtagatctg tagatttagc tatttttct ttacatttag ctggaatttc
421 ctctatttta ggtgctatta attttattac tacaattatc aatatacgat taaatagttt
481 atcatttgat caaatacctt tatttatttg agctgttggg attactgcat ttttactatt
541 attatcttta cctgtattag caggagctat tactatactt ttaacagatc gaaatcttaa
601 tacatctttt tttgatcctg ctggaggagg tgatccaatt ttatatcaac atttattttg
661 atttttt
```



Wildcards	
\w	Letters, numbers and _
.	Any character except \n \r
\d	Numerical digits
\t	Tab
\r	Return character. Also used as the generic end-of-line character in TextWrangler
\n	Line-feed character. Also used as the generic end-of-line character in Notepad++
\s	Space, tab, or end of line
[A-Z]	A single character of the ranges indicated in square brackets
[^A-Z]	A single character including all characters *not* in the brackets. Note that this will include \n unless otherwise specified, and may cause you to match across lines
\	Used to escape punctuation characters so they are searched for as themselves, not interpreted as wildcards or special symbols
\\	The \ symbol itself, escaped
Boundaries	
^	Match the start of the line, i.e., the position before the first character
$	Match the last position before the end-of-line character

Quantifiers, used in combination with characters and wildcards	
+	Look for the longest possible match of one or more occurrences of the character, wildcard, or bracketed character range immediately preceding. The match will extend as far as it can while still allowing the entire expression to match.
*	As above, matches as many of the previous character to occur, but allows for the character not to occur at all if the match still succeeds
?	Modifies greediness of + or * to match the shortest possible match instead of longest
{}	Specify a range of numbers to repeat the match of the previous character. For example: \d{2,4} matches between 2 and 4 digits in a row [AC]{4,} matches 4 or more of the letter A or C in a row
Capturing and replacing	
()	Capture the search results between the parentheses for use in the replacement term
\1 $1	Substitute the contents of the matched into the replacement term, in numerical order. Syntax depends on the text editor or language that you are using.