

Genome 540 Discussion

Jan. 29th 2026



GBFF file format

Header

```
LOCUS      U00096      4641652 bp    DNA      circular BCT 01-AUG-2014
DEFINITION Escherichia coli str. K-12 substr. MG1655, complete genome.
ACCESSION  U00096
VERSION    U00096.3
DBLINK     BioProject: PRJNA225
           BioSample: SAMN02604091
KEYWORDS   .
SOURCE     Escherichia coli str. K-12 substr. MG1655
  ORGANISM Escherichia coli str. K-12 substr. MG1655
           Bacteria; Proteobacteria; Gammaproteobacteria; Enterobacterales;
           Enterobacteriaceae; Escherichia.
REFERENCE  1 (bases 1 to 4641652)
  AUTHORS  Blattner,F.R., Plunkett,G. III, Bloch,C.A., Perna,N.T., Burland,V.,
           Riley,M., Collado-Vides,J., Glasner,J.D., Rode,C.K., Mayhew,G.F.,
           Gregor,J., Davis,N.W., Kirkpatrick,H.A., Goeden,M.A., Rose,D.J.,
           Mau,B. and Shao,Y.
  TITLE    The complete genome sequence of Escherichia coli K-12
  JOURNAL  Science 277 (5331), 1453-1462 (1997)
  PUBMED   9278503
REFERENCE  2 (bases 1 to 4641652)
  AUTHORS  Hayashi,K., Morooka,N., Yamamoto,Y., Fujita,K., Isono,K., Choi,S.,
           Ohtsubo,E., Baba,T., Wanner,B.L., Mori,H. and Horiuchi,T.
  TITLE    Highly accurate genome sequences of Escherichia coli K-12 strains
           MG1655 and W3110
  JOURNAL  Mol. Syst. Biol. 2, 2006 (2006)
  PUBMED   16738553
REFERENCE  3 (bases 1 to 4641652)
  AUTHORS  Riley,M., Abe,T., Arnaud,M.B., Berlyn,M.K., Blattner,F.R.,
           Chaudhuri,R.R., Glasner,J.D., Horiuchi,T., Keseler,I.M., Kosuge,T.,
           Mori,H., Perna,N.T., Plunkett,G. III, Rudd,K.E., Serres,M.H.,
           Thomas,G.H., Thomson,N.R., Wishart,D. and Wanner,B.L.
```

Features

```
FEATURES             Location/Qualifiers
     source            1..4641652
                        /organism="Escherichia coli str. K-12 substr. MG1655"
                        /mol_type="genomic DNA"
                        /strain="K-12"
                        /sub_strain="MG1655"
                        /db_xref="taxon:511145"
     gene              190..255
                        /gene="thrL"
                        /locus_tag="b0001"
                        /gene_synonym="ECK0001"
                        /gene_synonym="JW4367"
                        /db_xref="EcoGene:Eg11277"
     CDS               190..255
                        /gene="thrL"
                        /locus_tag="b0001"
                        /gene_synonym="ECK0001"
                        /gene_synonym="JW4367"
                        /function="leader; Amino acid biosynthesis: Threonine"
                        /note="G0 process: G0:0009088 - threonine biosynthetic
                        process"
                        /codon_start=1
                        /transl_table=11
                        /product="thr operon leader peptide"
                        /protein_id="AAC73112.1"
                        /db_xref="ASAP:ABE-00000006"
                        /db_xref="UniProtKB/Swiss-Prot:P0AD86"
                        /db_xref="EcoGene:Eg11277"
                        /translation="MKRISTITITITITITGNGAG"
```

Sequence

```
ORIGIN
1   agcttttcat tctgactgca acgggcaata tgtctctgtg tggattaaaa aaagagtgtc
61  tgatagcagc ttctgaactg gttacctgcc gtgagtaaat taaaatttta ttgacttagg
121 tcactaaata ctttaaccaa tataggcata gcgcacagac agataaaaat tacagagtac
181 acaacatcca tgaaacgcat tagcaccacc attaccacca ccataccatc taccacaggt
241 aacgggtcgg gctgacgcgt acaggaaaca cagaaaaaag cccgcacctg acagtgcggg
301 cttttttttt cgaccaaagg taacgaggta acaacatgac gagtgttgaa gticggcggt
361 acatcagttg caaatgcaga acgttttctg cgtgtgtccg atattctgga aagcaatgcc
421 aggcaggggc aggtggccac cgtcctctct gccccgccca aaatcaccaa ccacctggtg
481 gcgatgattg aaaaaacatc tagcggccag gatgctttac ccaatatcag cgatgccgaa
541 cgtatttttg ccgaactttt gacgggactc gccgccgccc agccggggtt cccgctggcg
601 caattgaaaa ctttcgtcga tcagggaattt gcccaataaa aacatgtcct gcattggcatt
661 agtttgttgg ggcagtgccc ggatagcate aacgctgcgc tgatttgcgc tggcgagaaa
721 atgctgatcg ccattatggc cggcgatta gaagcgcgcg gtcacaacgt tactgttatc
781 gatccggtcg aaaaactgct ggcagtgggg cattacctcg aatctaccgt cgattattgt
841 gagtccaccc gccgtattgc ggcaagccgc attccggctg atcacatggt gctgatggca
901 ggtttcacgc ccggtaatga aaaagcgcga cttgtgtgtc ttggacgcaa cggttcgcac
961 tactctgctg cgtgtctggc tgcctgttta cgcgccgatt gtgctgagat ttggacggac
1021 gttgacgggg tctatacctg cgaccgcgct cagggtgccc atgcgaggtt gttgaagtcg
1081 atgtcttacc aggaagcgat ggagctttcc tactctggcg ctaaaagtct tcacccccgc
1141 accattaccc ccacgcacca gttccagatc ccttgctgta ttaaaaatac cggaatcct
1201 caagcaccag gtacgctcat tggtgccagc cgtgatgaag acgaattacc ggtaagggac
1261 atttccaatc tgaataacat ggcaatgttc agcgtttctg gtccgggagt gaaagggatg
1321 gtccgcatag cggcgcgcgt ctttgacgac atgtcacgac cccgtatttc cgtgtgtctg
1381 attacgcaat catcttcgca atacagcatc agttttctgc ttccacaaaag cgactgtgtg
1441 cgagctgaac gggcaatgca ggaagagttc tactctggaac tgaagaagag cttactggag
1501 ccgtctgcag tgacggaacg gctggccatt atctcggtg taggtgatgg tatgcgcacc
1561 ttgcgtggga tctcggcgaa attctttgcc gcactggccc gcccaatat caacattgtc
```

Cliff Rostomily & Joe Min

GBFF file format

```
FEATURES
  source
    Location/Qualifiers
      1..1746380
      /organism="Streptococcus pyogenes"
      /mol_type="genomic DNA"
      /strain="NCTC12064"
      /serovar="Lancefield Group A"
      /isolation_source="not available: to be reported later"
      /culture_collection="NCTC:12064"
      /db_xref="taxon:1314"
      /chromosome="1"
      /country="United Kingdom: Telford"
      /collection_date="1900/1982"
  gene
    1..1356
    /gene="dnaA"
    /locus_tag="DQM35_RS00005"
    /old_locus_tag="NCTC12064_00001"
    /db_xref="GeneID:69899953"
  CDS
    1..1356
    /gene="dnaA"
    /locus_tag="DQM35_RS00005"
    /old_locus_tag="NCTC12064_00001"
    /inference="COORDINATES: similar to AA
    sequence:RefSeq:WP_012657571.1"
    /GO_function="GO:0003677 - DNA binding [Evidence IEA]"
    /GO_function="GO:0003688 - DNA replication origin binding
    [Evidence IEA]"
    /GO_function="GO:0005524 - ATP binding [Evidence IEA]"
    /GO_process="GO:0006270 - DNA replication initiation
    [Evidence IEA]"
    /GO_process="GO:0006275 - regulation of DNA replication
    [Evidence IEA]"
    /note="Derived by automated computational analysis using
    gene prediction method: Protein Homology."
    /codon_start=1
    /transl_table=11
    /product="chromosomal replication initiator protein DnaA"
    /protein_id="WP_002987659.1"
    /db_xref="GeneID:69899953"
    /translation="MTENEQIFWNRVLELAQSQLKQATYEFFVHDARLLKVKDHIATI
    YLDQMKELFWEKNLKDVLITAGFEVYNAQISVDYVFEEELMIEQNQTKINQPKQQAAL
    NSLPTVTSDLNSKYSFENFIQDENRWAVAASIAVANTPGTTYNPLFIWGGPGLGKTH
    LLNATIGNSVLLNPNAIKYITAENFINEFVIHRLDTMDLKEKFRNLDLLLIDDIQ
    SLAKKTLSGTQEEFNFTNALHNNKQIVLTSDRTPDHLNLEDRLVTRFKWGLTVNI
    TPDFETRVAILTNKIQEYNFIFPQDTIEYLAGQFDSNVRDLEGALKDISLVANFKQI
    DTITVDIAAEAIRARKQDGPKMTVIPIEEIQAQVGKFGYGVTVKEIKATKRTQNIVLAR
    QVAMFLAREMTDNLSPKIGKEFGGRDHSTVLHAYNKIKNMISQDESRLIEIETIKNKI
```

Gene: includes
exons, introns,
UTRs

CDS: respects
strand, excludes
introns.
(use this! first
annotated
nucleotide = TSS)

GBFF file format

```
gene      complement(1233609..1234707)
          /gene="prfB"
          /locus_tag="DQM35_RS06355"
          /old_locus_tag="NCTC12064_01257"
          /old_locus_tag="NCTC12064_01257"
CDS       complement(join(1233609..1234634,1234636..1234707))
          /gene="prfB"
          /locus_tag="DQM35_RS06355"
          /old_locus_tag="NCTC12064_01257"
          /inference="COORDINATES: similar to AA
          sequence:RefSeq:WP_011889043.1"
          /ribosomal_slippage
          /GO_function="GO:0003747 - translation release factor
          activity [Evidence IEA]"
          /GO_process="GO:0006415 - translational termination
          [Evidence IEA]"
          /note="programmed frameshift; Derived by automated
          computational analysis using gene prediction method:
          Protein Homology."
          /codon_start=1
          /transl_table=11
          /product="peptide chain release factor 2"
          /protein_id="WP_096466706.1"
          /db_xref="GeneID:69901161"
          /translation="MEVAEIRQKIVENKEKLTFRRLDLDRLEEEIALLENHMTPEP
          FWNNDNIAAQKTSQELNELKGKYDTFHNMQELSDTELLLEMLDEDDSLKEELEENLMQ
          LDKIMGAYEMTLLSEPYDHNNAILIHPGSGGTEAQDWGDLLLRMYTRFGNANGFKV
          EVLDYQAGDEAGIKSVTSLFEGPNAYGLLKSEMGVHRLVRISPFDSAKRRHTSFASVE
          VMPELDDTIEVEVRDDDIKMDTFRSGGAGGQNVNKVSTGVRLTHIPTGIVVSSTVDRT
          QYGNRDRAMKMLQAKLYQLEQEKKAQEVDAKKGDKKEITWGSQIRSYVFTPYTMVKDH
          RTNFELAQVDKVMGDGEINGFIDAYLKWRIED"
```

1233609..1234634

means

“the region (inclusive) between 1233609
and 1234634”

join(r1start..r1end, r2start..r2end) means
append the two regions

complement(start..end) means take the
reverse complement of the region

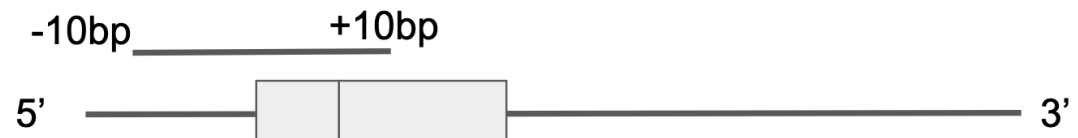
Examples:

Example: join(15..20,25..35)

15..20,25..35



join(15..20,25..35)



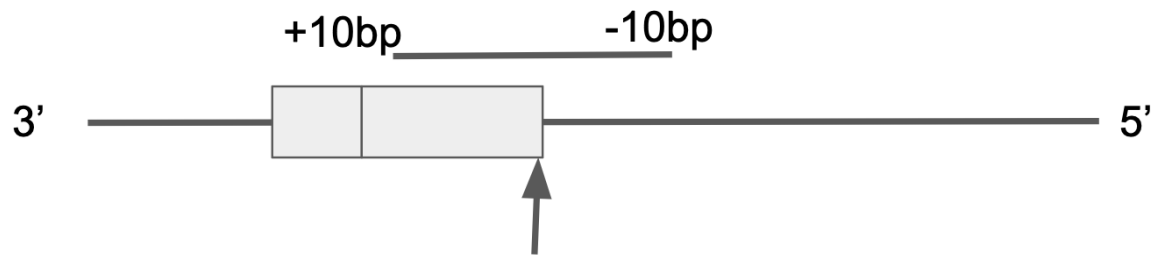
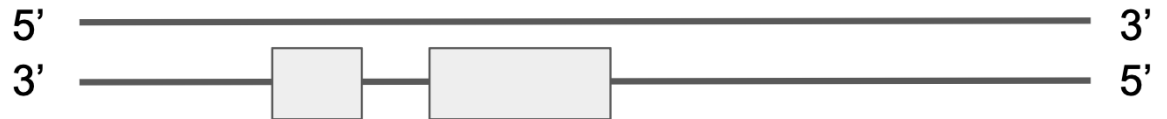
Examples:

Example: `complement(join(15..20,25..35))`

`15..20,25..35`

- Coordinates on + strand
- But take sequence on reverse complement

`join(15..20,25..35)`



TSS is at pos 35

One final gotcha:

```
test1.gbff.txt
LOCUS      U00096          4641652 bp    DNA      circular BCT 01-AUG-2014
DEFINITION Escherichia coli str. K-12 substr. MG1655, complete genome.
ACCESSION  U00096
VERSION    U00096.3
DBLINK     BioProject: PRJNA225
           BioSample: SAMN02604091
KEYWORDS   .
SOURCE     Escherichia coli str. K-12 substr. MG1655
  ORGANISM Escherichia coli str. K-12 substr. MG1655
           Bacteria; Proteobacteria; Gammaproteobacteria; Enterobacterales;
           Enterobacteriaceae; Escherichia.
REFERENCE  1 (bases 1 to 4641652)
  AUTHORS  Blattner,F.R., Plunkett,G. III, Bloch,C.A., Perna,N.T., Burland,V.,
           Riley,M., Collado-Vides,J., Glasner,J.D., Rode,C.K., Mayhew,G.F.,
           Gregor,J., Davis,N.W., Kirkpatrick,H.A., Goeden,M.A., Rose,D.J.,
           Mau,B. and Shao,Y.
  TITLE    The complete genome sequence of Escherichia coli K-12
  JOURNAL   Science 277 (5331), 1453-1462 (1997)
  PUBMED    9278503
```

```
test2.gbff.txt
LOCUS      NC_004331      2895605 bp    DNA      linear    INV 29-JUL-2010
DEFINITION Plasmodium falciparum 3D7 chromosome 13.
ACCESSION  NC_004331
VERSION    NC_004331.2  GI:296005645
DBLINK     Project: 148
KEYWORDS   complete genome.
SOURCE     Plasmodium falciparum 3D7
  ORGANISM Plasmodium falciparum 3D7
           Eukaryota; Alveolata; Apicomplexa; Aconoidasida; Haemosporida;
           Plasmodium; Plasmodium (Laverania).
REFERENCE  1 (bases 1 to 2895605)
  AUTHORS  Harris,B., Lennard,N., Clark,L., Line,A., Barron,A., Corton,C.,
           Berriman,M., Pain,A., Hall,N., Atkin,R., Chillingworth,C.,
           Doggett,J., Ormond,D., Sanders,M., Hayes,R., Hall,S., Quail,M. and
           Barrell,B.
  JOURNAL   Unpublished
REFERENCE  2 (bases 1 to 2895605)
  CONSRTM   NCBI Genome Project
  TITLE     Direct Submission
```

One final gotcha:

```
LOCUS      NZ_LS483338          1746380 bp    DNA      circular CON 25-DEC-2022
DEFINITION Streptococcus pyogenes strain NCTC12064 chromosome 1, complete
            sequence.
ACCESSION  NZ_LS483338
VERSION    NZ_LS483338.1
DBLINK     BioProject: PRJNA224116
            BioSample: SAMEA3594357
            Assembly: GCF_900475035.1
KEYWORDS   RefSeq.
SOURCE     Streptococcus pyogenes
ORGANISM   Streptococcus pyogenes
            Bacteria; Firmicutes; Bacilli; Lactobacillales; Streptococcaceae;
            Streptococcus.
```

```
CDS        /db_xref="GeneID:69899953"
            1..1356
            /gene="dnaA"
            /locus_tag="DQM35_RS00005"
            /old_locus_tag="NCTC12064_00001"
            /inference="COORDINATES: similar to AA
            sequence:RefSeq:WP_012657571.1"
            /GO_function="GO:0003677 - DNA binding [Evidence IEA]"
            /GO_function="GO:0003688 - DNA replication origin binding
            [Evidence IEA]"
            /GO_function="GO:0005524 - ATP binding [Evidence IEA]"
            /GO_process="GO:0006270 - DNA replication initiation
            [Evidence IEA]"
```

Another Gotcha

Some CDSs are on multiple lines!

Linked List...?



regex: Regular Expressions

Popular way of describing general patterns in text

May have seen: `grep`

from the editor `ed`:

`g/re/p` (global / regular expression / print)

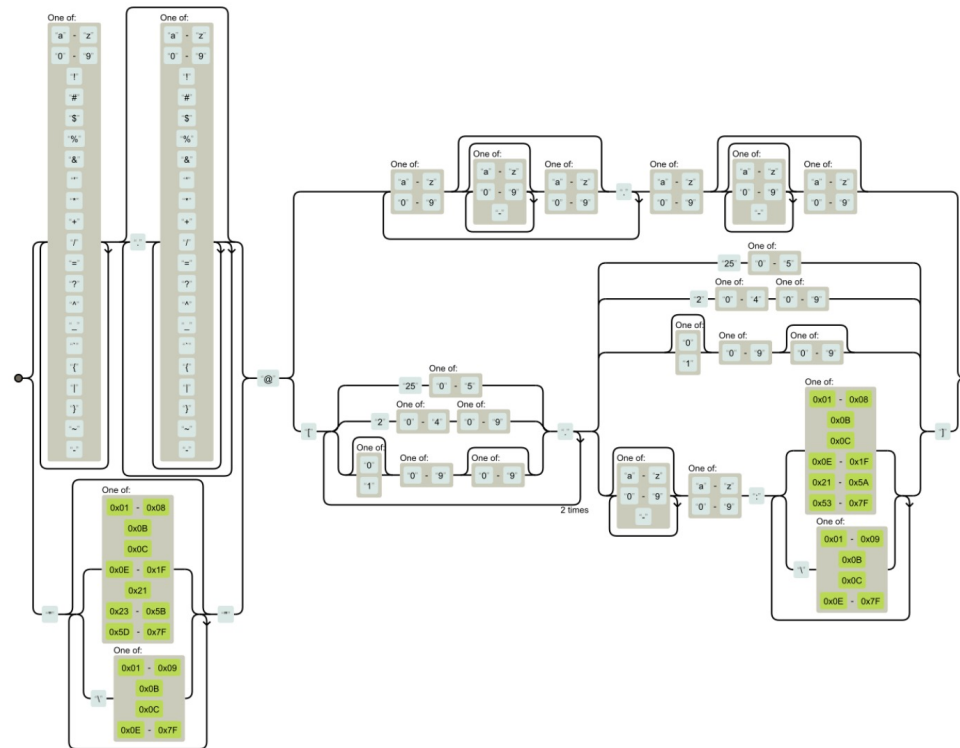
Every language will have a regex engine

regex: Regular Expressions

```
(?:[a-z0-9!#$%&'*/+=?^_`{|}~-]+(?:\.(?:[a-z0-9!#$%&'*/+=?^_`{|}~-]+)*|"(?:[\x01-\x08\x0b\x0c\x0e-\x1f\x21\x23-\x5b\x5d-\x7f]|\\[\x01-\x09\x0b\x0c\x0e-\x7f])*")@(?:(?:[a-z0-9](?:[a-z0-9-]*[a-z0-9])?\.|)(?:[a-z0-9](?:[a-z0-9-]*[a-z0-9])?)|(?:(?:25[0-5]|2[0-4][0-9]|01[0-9]?[0-9]?[0-9])\.){3}(?:25[0-5]|2[0-4][0-9]|01[0-9]?[0-9]?[0-9])?[a-z0-9-]*[a-z0-9]:(?:[\x01-\x08\x0b\x0c\x0e-\x1f\x21-\x5a\x53-\x7f]|\\[\x01-\x09\x0b\x0c\x0e-\x7f])+)?)\])
```

<https://emailregex.com/>

regex: Regular Expressions



<https://emailregex.com/>

regex: Regular Expressions

Easiest case: just matching text:

hello.txt:

```
1 Hello there
2 General kenobi
3 you are a bold one
4 kill him!
5 [droids get crushed]
6 back away! I will deal with this Jedi slime myself
7 your move
8 you fool
9 I've been trained in your Jedi arts by Count Dooku
10 attack kenobi
11
```

```
hex ~/🐙 grep "kenobi" hello.txt
General kenobi
attack kenobi
hex ~/🐙
```

regex: Regular Expressions

❑ The “.” Metacharacter:

- The period (.) represents a single wildcard character.
- It will match on any single character except for a newline.
- If you want to treat this metacharacter as a period character rather than a wildcard, precede it with a backslash (\.) to escape its special meaning.

Example 3-1. frost.txt

```
1 Two roads diverged in a yellow wood,  
2 And sorry I could not travel both  
3 And be one traveler, long I stood  
4 And looked down one as far as I could  
5 To where it bent in the undergrowth;  
6  
7 Excerpt from The Road Not Taken by Robert Frost
```

```
$ grep 'T.o' frost.txt
```

```
1 Two roads diverged in a yellow wood,
```

- If we try to match on the pattern T.o, the first line of the frost.txt file is returned because it contains the word Two
- Regex patterns are also **case sensitive**, which is why line 3 of the file is not returned even though it contains the string **too**

regex: Regular Expressions

❑ The “?” Metacharacter:

- The question mark (?) character makes any item that **precedes** it **optional**.
- It matches it zero or one time.

```
$ egrep 'T.?o' frost.txt
```

```
1  Two roads diverged in a yellow wood,  
5  To where it bent in the undergrowth;
```

- ❑ This pattern will match on any **three-character** sequence that begins with T and ends with o as well as the **two-character** sequence **To**.

regex: Regular Expressions

❑ The “*” Metacharacter:

- The asterisk (*) is a special character that **matches** the **preceding** item **zero or more times**.
- It is **similar** to **?**, the main difference being that the previous item may appear **more than once**.

```
$ grep 'T.*o' frost.txt
```

```
1  Two roads diverged in a yellow wood,  
5  To where it bent in the undergrowth;  
7  Excerpt from The Road Not Taken by Robert Frost
```

- ❑ The .* in the preceding pattern allows **any number** of any character to appear between the **T** and **o**.
- ❑ Thus, the last line also matches because it contains the pattern **The Ro.**

regex: Regular Expressions

❑ The “+” Metacharacter:

- The plus sign (+) metacharacter is the same as the * except it requires the **preceding** item to appear at **least once**.

```
$ egrep 'T.+o' frost.txt
```

```
1    Two roads diverged in a yellow wood,  
5    To where it bent in the undergrowth;  
7 Excerpt from The Road Not Taken by Robert Frost
```

- ❑ The preceding pattern specifies one or more of any character to appear in between the T and o.
- ❑ The first line of text matches because of Two — the w is one character between the T and the o.
- ❑ The second line doesn't match the To, as in the previous example; rather, the pattern matches a much larger string — all the way to the o in undergrowth.
- ❑ The last line also matches because it contains the pattern The Ro.

regex: Regular Expressions

- ❑ We can use **parentheses** to group characters.
- ❑ Among other things, this allows us to **treat** the characters appearing inside the parentheses as **a single item** that we can later reference.

```
$ egrep 'And be one (stranger|traveler), long I stood' frost.txt  
3    And be one traveler, long I stood
```

- ❑ Here, we use parentheses and the Boolean OR operator (|) to create a pattern that will match on line 3.
- ❑ Line 3 as written has the word traveler in it, but this pattern would match even if traveler was replaced by the word stranger.

regex: Regular Expressions

- ❑ **Table 3-1** provides a list of common examples when using character classes and ranges.

Table 3-1. Regex character ranges

Example	Meaning
[abc]	Match only the character <i>a</i> or <i>b</i> or <i>c</i>
[1-5]	Match on digits in the range 1 to 5
[a-zA-Z]	Match any lowercase or uppercase <i>a</i> to <i>z</i>
[0-9 +-* /]	Match on numbers or these four mathematical symbols
[0-9a-fA-F]	Match a hexadecimal digit

- ❑ **Be careful** when defining a range for digits; the range can at most go from 0 to 9. For example, the pattern [1-475] does not match on numbers between 1 and 475; it matches on any one of the digits (characters) in the range 1–4 or the character 7 or the character 5.

regex: Regular Expressions

- ❑ There are also predefined character classes known as **shortcuts**.
- ❑ These can be used to indicate common character classes such as numbers or letters.

Table 3-2. Regex shortcuts

Shortcut	Meaning
\s	Whitespace
\S	Not whitespace
\d	Digit
\D	Not digit
\w	Word
\W	Not word
\x	Hexadecimal number (e.g., 0x5F)

<https://www.slideshare.net/slideshow/chapter-3-introduction-to-regular-expression/266815459>

regex: Regular Expressions

- ❑ You can use **anchors** to specify that a pattern must exist at the beginning or the end of a string.
- ❑ The caret (^) character is used to anchor a pattern to the beginning of a string.
- ❑ **For example**, **^[1-5]** means that a matching string must start with one of the digits 1 through 5, as the first character on the line.
- ❑ The **\$** character is used to anchor a pattern to the **end** of a string or line.
- ❑ **For example**, **[1-5]\$** means that a string must end with one of the digits 1 through 5.
- ❑ In addition, you can use **\b** to identify a word boundary (i.e., a space).
 - ❑ The pattern **\b[1-5]\b** will match on any of the digits 1 through 5, where the digit appears as its own word.

regex: Regular Expressions



4395



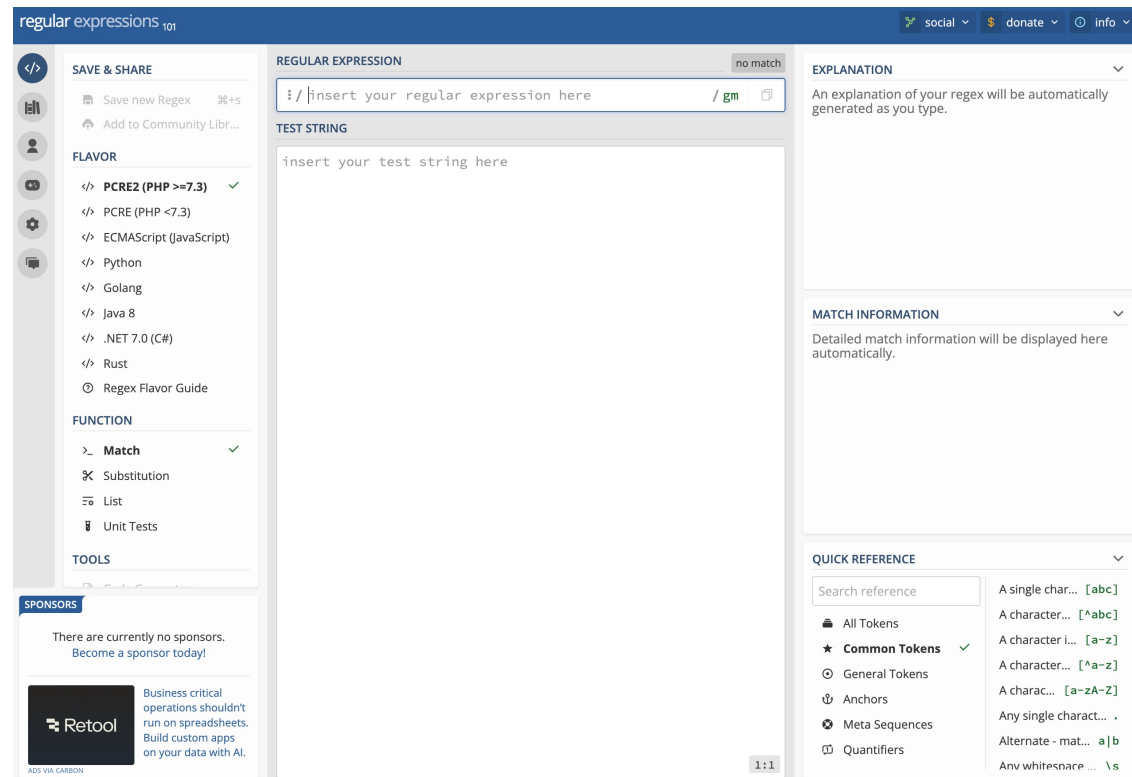
Locked. There are [disputes about this answer's content](#) being resolved at this time. It is not currently accepting new interactions.



You can't parse [X]HTML with regex. Because HTML can't be parsed by regex. Regex is not a tool that can be used to correctly parse HTML. As I have answered in [HTML-and-regex](#) questions here so many times before, the use of regex will not allow you to consume HTML. Regular expressions are a tool that is insufficiently sophisticated to understand the constructs employed by HTML. HTML is not a regular language and hence cannot be parsed by regular expressions. Regex queries are not equipped to break down HTML into its meaningful parts. so many times but it is not getting to me. Even enhanced irregular regular expressions as used by Perl are not up to the task of parsing HTML. You will never make me crack. HTML is a language of sufficient complexity that it cannot be parsed by regular expressions. Even Jon Skeet cannot parse HTML using regular expressions. Every time you attempt to parse HTML with regular expressions, the unholy child weeps the blood of virgins, and Russian hackers

<https://stackoverflow.com/questions/1732348/regex-match-open-tags-except-xhtml-self-contained-tags>

regex: Demo?



<https://regex101.com/>