A Quick Guide To PERL Regular Expressions

This is a Quick reference Guide for PERL regular expressions (also known as regexps or regexes). These tools are used to describe text as “motifs” or “patterns” for matching, quoting, substituting or transliterating. Each programming language (Perl, C, Java, Python...) define its own regular expressions although the syntax might differ from details to extensive changes. In this guide we will concentrate on the Perl regex syntax, we assume that the reader has some preliminary knowledge of Perl programming.

Perl uses a Traditional Nondeterministic Finite Automata (NFA) match engine. This means that it will compare each element of the motif to the input string, keeping track of the positions. The engine choose the first leftmost match after greedy (i.e., longest possible match) quantifiers have matched.

References

For more information on Perl regexps and other syntaxes you can refer to O’Reilly’s book “Mastering Regular Expressions”.

Examples:
The following sentence will be used in all our examples:
The ID sp:UBP5_RAT is similar to the rabit AC tr:Q12345

Motif finding: match operator m/

Example: correct typo for the word rabbit

Example: reuse of a precompiled regexp

Search&Replace: substitution operator s///

Example: correct typo for the word rabbit

Example: reuse of a precompiled regexp

POSIX Character class

Special characters

Repetitions

Anchors

Capture & Grouping

Example: match any database code in the list

Options

e evaluate REPLACE as an expression
g global matches (matches all occurrences)
i case insensitive
m multiline, allow "*" and "$" to match with (\n)
o compile MOTIF only once
s single line, dot "." matches new-line (\n)
x ignore whitespace and allow comments "#" in MOTIF

Quoting: quote and compile operator qr//

Examples: reuse of a precompiled regexp

Character classes

Capture & Grouping

Example: match any database code in the list

$expr = m/(sp:tr:rs:)/g;

will match:

The ID sp:UBP5_RAT is similar to the rabit AC tr:Q12345

Options

cg continue after a failure in /g
g global matches (matches all occurrences)
i case insensitive
m multiline, allow "*" and "$" to match with (\n)
o compile MOTIF only once
s single line, dot "." matches new-line (\n)
x ignore whitespace and allow comments "#" in MOTIF

Character classes

[...] Match any one character of a class
[.\[] Match any one character not in the bracket
. Match any character (except newline [\n]) in non single-line mode (/s)
\d Any digit. Equivalent to [0..9] or [1:digit:]
\D Any non-digit.
\s Any whitespace. [ \t\n\f\r\v\n\f\r\v] or [[space:]]
\S Any non-whitespace.
\w Any word character. [a-zA-Z0-9-_] or [[alnum:]]
Back reference. Match the same as the captured
group number \( n \) that was previously matched in
the same MOTIF.

Substring of captured group \( n \)

Example: match several instances with back reference
$ex =~ m/(the).+\1/1;
will match:
The ID sp:UBP5_RAT is similar to the rabbit AC tr:Q12345

Example: rename any tr:AC to trembl_AC= using a capture
$ex =~ s/tr:(\[[[:alnum:]]\]{6})/trembl_AC=$1/gi;
will match:
The ID sp:UBP5_RAT is similar to the rabbit AC trembl_AC=Q12345

Text-span modifiers
\Q Quote following metacharacters until \E or end of
motif (allow the use of scalars in regexp)
\u Force next character to uppercase
\l Force next character to lowercase
\U Force all following characters to uppercase
\L Force all following characters to lowercase
\E End a span started with \Q, \U or \L

Extended Regexp
(?#...) Substring “…” is a comment
(?=...) Positive lookahead. Match if exists next match
(e.g., allow overlapping matches in global mode)
(?1...) Negative lookahead. Match if no next match
(?<=...) Positive lookbehind. Fixed length only.
(?<1...) Negative lookbehind. Fixed length only.
(?1mxx) Modify matching options

Transliteration: translate operator tr///
EXPR =~ tr/SEARCHLIST/REPLACELIST/cds

Transliteration is not - and does not use - a regular expression,
but it is frequently associated with the regexp in PERL. Thus
we decided to include it in this guide.

Example: reverse and complement a DNA sequence
$DNA = AAATATTTCATGTCATACAT;
$revcom = reverse $DNA;
$revcom =~ tr/ACGT TGCA/;

The transliteration will produce the following:
print($DNA);
AAATATTTCATGTCATACAT
print($revcom);
TGCAATCGTTATCACAT

Options
c complement REPLACELIST
d delete non-replaced characters
s single replace of duplicated characters

UniCode matches
Perl 5.8 supports UniCode 3.2. However it would be too
long to describe all the properties in details here. For more
information see “Mastering Regular Expressions”.

\p{PROP} Matches a UniCode property
\P{PROP} Matches anything but a Uni Code property

This document was written and designed by Laurent Falquet
and Vassilios Ioannidis from the Swiss EMBnet node and being
distributed by P&PR Publications Committee of EMBnet.

EMBnet - European Molecular Biology Network - is a
bioinformatics support network of bioinformatics support
centers situated primarily in Europe. Most countries have a
national node which can provide training courses and other
forms of help for users of bioinformatics software.

You can find information about your national node from the
EMBnet site:
http://www.embnet.org/

A Quick Guide To PERL Regular Expressions
Second edition © 2006