

Regular Expressions (RegEx)

Regular expression is like another language

- What is a regular expression?
- Literal (or normal characters)
 - Alphanumeric
 - abc...ABC...0123...
 - Punctuation
 - `-_.,:;=()/+ *%&{}[]? !$'^|\<>"@#`
- Just like languages Regular expressions also have dialects
 - awk, egrep, Emacs, grep, Perl, POSIX, Tcl, PROSITE

Why use a regular expression?

ALDVANRPMKPEMFAAHRAKTLAELRKRKLEGVVLIYGFP
EPTRAHCDFEPVFRQESCFYWLTGVNEADCAYFLDIETGKEIL
FYDPDIPQAYIIWFGELATIDDIKKKYGFEDVRLMPKIQETLAEY
KLKKIHTLPETCILKGYVAVKDKNEFIDVVGELRQIKDDDEM
LIQYACDVNSFAVRDTFKKVHPKMWEHQVEANLIKHYVDYY
CRCFAFSTIVCSGENCSILHYHHNNKFIEDGELILIDTGCEYNC
AADNTRTIPANGKFSPDQRAVYQAVLDCHNYVVAHAKPGV
WPDLAYDSAKVMAAGLLKLGLFQNGTVDEIVDAGALAVFY
PHGLGHGMGIDCHEIAGWPRGTCRGKKPHHSFVRFGRTE
KGVVITNEPGCYFIRPSYNAAFADPEKSKYINKEVCERLRKTV
GGVRIEDDLLITEDGCKVLSNIPKEIDEIEAFMAKKESKL

FGCV . . . WKAQLLNEY VAVK FPIQ . DKQSWQNEY EVYS LPGM . K
YGEV . . . WRGSWQGEN VAVK FSSR . DEKSWFRET ELYNTVML . R
FGKV . . . YRAFWIGDE VAVK ARHDP . DEDISQTIENVR QEAKLFAML . K
FGTV . . . YKKGKWHGD VAVK LKVV . DPTPEQFQAFRN EVAVLRKT . R
SGTV . . . YKGVLEDDRH VAVK LENV . RQGKEVFQA ELSVIGRI . N

VAVK

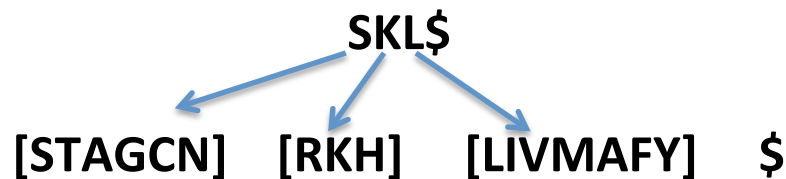
Why use a regular expression?

ALDVANRPMPKPEMFAAHRAKTLAELRKRKLEGVVLIYGFP
EPTRAHCDFEPVFRQESCFYWLTGVNEADCAYFLDIETGKEIL
FYDPDIPQAYIIWFGELATIDDIKKKYGFEDVRLMPKIQETLAEY
KLKKIHTLPETCILKGY**VAVK**DKNEFIDVVGELRQIKDDDEM
LIQYACDVNSFAVRDTFKKVHPKMWEHQVEANLIKHYVDYY
CRCFAFSTIVCSGENCSILHYHHNNKFIEDGELILIDTGCEYNC
AADNTRTIPANGKFSPDQRAVYQAVLDCHNYVVAHAKPGV
WPDLAYDSAKVMAAGLLKLGLFQNGTVDEIVDAGALAVFY
PHGLGHGMGIDCHEIAGWPRGTCRGKKPHHSFVRFGRLE
KGVVITNEPGCYFIRPSYNAAFADPEKSKYINKEVCERLRKTV
GGVRIEDDLLITEDGCKVLSNIPKEIDEIEAFMAKKESKL

Regular expression = vavk

Why use a regular expression?

ALDVANRPMPKPEMFAAHRAKTLAELRKRKLEGVVLIYGFP
EPTRAHCDFEPVFRQESCFYWLTGVNEADCAYFLDIETGKEIL
FYDPDIPQAYIIWFGELATIDDIKKKYGFEDVRLMPKIQETLAEY
KLKKIHTLPETCILKGYVAVKDKNEFIDVVGELRQIKDDDEM
LIQYACDVNSFAVRDTFKKVHPKMWEHQVEANLIKHYVDYY
CRCFAFSTIVCSGENCSILHYHHNNKFIEDGELILIDTGCEYNC
AADNTRTIPANGKFSPDQRAVYQAVLDCHNYVVAHAKPGV
WPDLAYDSAKVMAAGLLKLGLFQNGTVDEIVDAGALAVFY
PHGLGHGMGIDCHEIAGWPRGTCRGKKPHHSFVRFGRTE
KGVVITNEPGCYFIRPSYNAAFADPEKSKYINKEVCERLRKTV
GGVRIEDDLLITEDGCKVLSNIPKEIDEIEAFMAKKE**SKL**



- **MLST**DNVANRPMKPEMF....
- Text: The sequence must start with an **methionine**, followed by any amino acid, followed by a **serine** or a **threonine**, two times, followed by any amino acid or nothing, followed by any amino acid except a **valine**.
- Regex: **$\text{^M.[ST]\{2\}.?[^V]}$**
Or: **$\text{^M.[ST]\{2\}.{0,1}[^V]}$**