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?

To find a pattern

MALDVANRPMPKPEMFAAHRAKTLAELRKRKLEGVVLIYGFPE
PTRAHCDFEPVFRQESCFYWLTGVNEADCAYFLDIETGKEILFYP
DIPQAYIIWFGELATIDDIQQQQQGFEDVRLMPKIQETLAEYKL
KKIHTLPETCILKGYVAVKDKNEFIDVVGELRQIKDDDEMVLIQY
ACDVNSFAVRDTFKKVHPKMWEHQVEANLIKHYVDYYCRCFA
FSTIVCSGENCSILHYHHNNKFIEDGELILIDTGCEYNCAADNTR
TIPANGKFSPQQQQQRAVYQAVVAVKLDCHNYVVAHAKPGV
WPDLAYDSAKVMAAGLLKLGLFQNGTVDEIVDAGALAVFYPH
GLGHGMGIDCHEIAHRAKGWPRGTCRGKKPHHSFVRFGRTLE
KGVVITNEPGCYFIRPSYNAAFADPEKSKYINKEVCERLRKTVGG
VRIEDDLLITEDGCKVLSNIPKEIHRAKDEIEAFMAKKESKL

Why use a regular expression?

To find a pattern

MALDVANRPMPKPEMFAAHRAKTLAELRKRKLEGVVLIYGFPE
PTRDRINKFEPVFRQESCFYWLTGVNEADCAYFLDIETGKEILFY
PDIPQAYIIWFGELATIDDI**QQQQQ**GFEDVRLMPKIQETLAEYK
LKKIHTLPETCILKGYVAVKDKNEFIDVVGELRQIKDDDEMVLIQ
YACDVNSFAVRDTFKKVHPKMWEHQVMILKHYVDYYCRCFAF
STIVCSGENCSILHYHHNNKFIEDGELILIDTGCEYNCAADNTRTI
PANGKFSP**QQQQQ**RAVYQAVVAVKLDCHNYVVAHAKPGVW
PDLAYDSAKVMAAGLLKLGLFQNGTVDEIVDAGALAVFYPHGL
GHGMGIDCHEIAHRAKGWPRGTCRGKKPHHSFVRFGRTLEKG
VVITNEPGCYFIRPSYNAAFADPEKSKYINKEVCERLRKTVGGVR
IEDDLLITEDGCKVLSNIPKEIHRAKDEIEAFMAKKESKL

Why use a regular expression?

To find a pattern

MALDVANRPMPKPEMFAAHRAKTLAEL**RKRK**LEGVVLIYGFP
EPTRDRINKEPVFRQESCFYWLTGVNEADCAYFLDIETGKEILFY
PDIPQAYIIWFGELATIDDI**QQQQQ**GFEDVRLMPKIQETLAEYK
LKKIHTL**RKRK**ILKGYVAVKDKNEFIDVVGELRQIKDDDEMVLIQ
YACDVNSFAVRDTFKKVHPKMWEHQVMILKHYVDYYCRCFAF
STIVCSGENCSILHYHHNNKFIEDGELILIDTGCEYNCAADNTRTI
PANGKFSP**QQQQQ**RAVYQAVVAVKLDCHNYVVAHAKPGVW
PDLAYDSAKVMAAGLLKLGLFQNGTVDEIVDAGALAVFYPHGL
GHGMGIDCHEIAHRAKGWPRGTCRGKKPHHSFVRFGRTLEKG
VVITNEPGCYFIRPSYNAAFADPEKSKY**RKRK**VCERLRKTVGGV
RIEDDLLITEDGCKVLSNIPKEIHRAKDEIEAFMAKKESKL

Why use a regular expression?

To find a pattern

MALDVANRPMPKPEMFAAHRAKTLAEL**RKRK**LEGVVLIYGFP
EPTR**DRINK**EPVFRQESCFYWLTGVNEADCAYFLDIETGKEILFY
PDIPQAYIIWFGELATIDDI**QQQQQ**GFEDVRLMPKIQETLAEYK
LKKIHTL**RKRK**ILKGYVAVKDKNEFIDVVGELRQIKDDDEMVLIQ
YACDVNSFAVRDTFKKVHPKMWEHQV**MILK**HYVDYYCRCFAF
STIVCSGENCSILHYHHNNKFIEDGELILIDTGCEYNCAADNTRTI
PANGKFSP**QQQQQ**RAVYQAVVAVKLDCHNYVVAHAKPGVW
PDLAYDSAKVMAAGLLKLGLFQNGTVDEIVDAGALAVFYPHGL
GHGMGIDCHEIAHRAKGWPRGTCRGKKPHHSFVRFGRTLEKG
VVITNEPGCYFIRPSYNAAFADPEKSKY**RKRK**VCERLRKTVGGV
RIEDDLLITEDGCKVLSNIPKEIHRAKDEIEAFMAKKESKL


VAVK

Why use a regular expression?

To find a pattern

MALDVANRPMPKPEMFAAHRAKTLAELRKRKLEGVVLIYGFPE
PTRDRINKEPVFRQESCFYWLTGVNEADCAYFLDIETGKEILFYP
DIPQAYIIWFGELATIDDIQQQQQGFEDVRLMPKIQETLAEYKL
KKIHTLRKRKILKGY**VAVK**DKNEFIDVVGELRQIKDDDEMVLIQY
ACDVNSFAVRDTFKKVHPKMWEHQVMILKHYVDYYCRCFAFS
TIVCSGENCSILHYHHNNKFIEDGELILIDTGCEYNCAADNTRTIP
ANGKFSPQQQQQRAVYQAV**VAVK**LDCHNYVVAHAKPGVWP
DLAYDSAKVMAAGLLKLGLFQNGTVDEIVDAGALAVFYPHGLG
HGMGIDCHEIAHRAKGWPRGTCRGKKPHHSFVRFGRTLEKGV
VITNEPGCYFIRPSYNAAFADPEKSKYRKRKVCERLRKTVGGVRI
EDDLLITEDGCKVLSNIPKEIHRAKDEIEAFMAKKESKL

Why use a regular expression?

MALDVANRPMPKPEMFAAHRAKTLAELRKRKLEGVVLIYGFPE
PTRDRINKEPVFRQESCFYWLTGVNEADCAYFLDIETGKEILFYP
DIPQAYIIWFGELATIDDIQQQQQGFEDVRLMPKIQETLAEYKL
KKIHTLRKRKILKGYVAVKDKNEFIDVVGELRQIKDDDEMVLIQY
ACDVNSFAVRDTFKKVHPKMWEHQVMILKHYVDYYCRCFAFS
TIVCSGENCSILHYHHNNKFIEDGELILIDTGCEYNCAADNTRTIP
ANGKFSPQQQQQRAVYQAVVAVKLDCHNYVVAHAKPGVWP
DLAYDSAKVMAAGLLKLGLFQNGTVDEIVDAGALAVFYPHGLG
HGMGIDCHEIAHRAKGWPRGTCRGKKPHHSFVRFGRTLEKGV
VITNEPGCYFIRPSYNAAFADPEKSKYRKRKVCERLRKTVGGVRI
EDDLLITEDGCKVLSNIPKEIHRAKDEIEAFMAK**KDEL**

KDEL$

[HK]DEL$

- MLSTDNVANRPMPKPEMF….

- 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: ^M.[ST]{2}.?[^V]

Useful RegEx help

- https://regex101.com
- https://regexr.com
- https://www.regextester.com
- https://medium.com/factory-mind/regex-tutorial-a-simple-cheatsheet-by-examples-649dc1c3f285

The YXXΦ (Y=tyrosine, X=any amino acid, Φ=bulky hydrophobic [phenylalanine, tyrosine , threonine]) motif is conserved in many eukaryotic membrane proteins that are recognized by adaptor proteins for sorting in the endosomal/lysosomal pathway. This motif is typically located in the c-terminal end of the protein. ****Note: do not look for the Φ symbol on your keyboard – this will not work. Rather you should use the amino acid symbols.*

Use the "protein motif pattern" search to find all proteins that contain this motif anywhere in the terminal 10 amino acids of proteins. (hint: for your regular expression, remember that you want the first amino acid to be a tyrosine, followed any two amino acids, followed by any bulky hydrophobic amino acid (phenylalanine, tyrosine, threonine).