

# *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?*

MALDVANRPMPKPEMFAAHRAKTLAELRKRKLEGVVLIYGFPE  
PTRDRINKEPVFRQESCFYWLTGVNEADCAYFLDIETGKEILFYP  
DIPQAYIIWFGELATIDDIQQQQQGFEDVRLMPKIQETLAEYKL  
KKIHTLRKRKILKGYVAVKDKNEFIDVVGELRQIKDDDEMVLIQY  
ACDVNSFAVRDTFKVHPKMWEHQVMILKHYVDYYCRCFAFS  
TIVCSGENCSILHYHHNNKFIEDGELILIDTGCEYNCAADNTRTIP  
ANGKFSPQQQQQRAVYQAVVAVKLDCHNYVVAHAKPGVWP  
DLAYDSA KVMAAGLLKLGLFQNGTVDEIVDAGALAVFYPHGLG  
HGMGIDCHEIAHRAKGWPRGTCRGKKPHHSFVRFGRTLEKGV  
VITNEPGCYFIRPSYNAAFADPEKSKYRKRKVCERLRKTGGVRI  
EDDLLITEDGCKVLSNIPKEIHRAKDEIEAFMAKKESKL

# *Why use a regular expression?*

To find a pattern

MALDVANRPMPKPEMFAAHRAKTLAELRKRKLEGVVLIYGFPE  
PTRDRINKEPVFRQESCFYWLTGVNEADCAYFLDIETGKEILFYP  
DIPQAYIIWFGELATIDDIQQQQQGFEDVRLMPKIQETLAEYKL  
KKIHTLRKRKILKGYVAVKDKNEFIDVVGELRQIKDDDEMVLIQY  
ACDVNSFAVRDTFKVHPKMWEHQVMILKHYVDYYCRCFAFS  
TIVCSGENCSILHYHHNNKFIEDGELILIDTGCEYNCAADNTRTIP  
ANGKFSPQQQQQRAVYQAVVAVKLDCHNYVVAHAKPGVWP  
DLAYDSA KVMAAGLLKLGLFQNGTVDEIVDAGALAVFYPHGLG  
HGMGIDCHEIAHRAKGWPRGTCRGKKPHHSFVRFGRTLEKGV  
VITNEPGCYFIRPSYNAAFADPEKSKYRKRKVCERLRKTGGVRI  
EDDLLITEDGCKVLSNIPKEIHRAKDEIEAFMAKKESKL

# *Why use a regular expression?*

To find a pattern

MALDVANRPMPKPEMFAAHRAKTLAELRKRKLEGVVLIYGFPE  
PTRDRINKEPVFRQESCFYWLTGVNEADCAYFLDIETGKEILFYP  
DIPQAYIIWFGELATIDDIQQQQQGFEDVRLMPKIQETLAEYKL  
KKIHTLRKRKILKGYVAVKDKNEFIDVVGELRQIKDDDEMVLIQY  
ACDVNSFAVRDTFKVHPKMWEHQVMILKHYVDYYCRCFAFS  
TIVCSGENCSILHYHHNNKFIEDGELILIDTGCEYNCAADNTRTIP  
ANGKFSPQQQQQRAVYQAVVAVKLDCHNYVVAHAKPGVWP  
DLAYDSAKVMAAGLLKLGLFQNGTVDEIVDAGALAVFYPHGLG  
HGMGIDCHEIAHRAKGWPRGTCRGKKPHHSFVRFGRTLEKGV  
VITNEPGCYFIRPSYNAAFADPEKSKYRKRKVCERLRKTGGVRI  
EDDLLITEDGCKVLSNIPKEIHRAKDEIEAFMAKKESKL

# *Why use a regular expression?*

To find a pattern

MALDVANRPMPKPEMFAAHRAKTLAELRKRKLEGVVLIYGFP  
EPTRDRINKEPVFRQESCFYWLTGVNEADCAYFLDIETGKEILFY  
PDIPQAYIIWFGELATIDDIQQQQQGFEDVRLMPKIQETLAEYK  
LKKIHTLRKRKIJKGYVAVKDKNEFIDVVGELRQIKDDDEMVLIQ  
YACDVNSFAVRDTFKVHPKMWEHQVMILKHVDYYCRCFAF  
STIVCSGENCSILHYHHNNKFIEDGELILIDTGCEYNCAADNRTI  
PANGKFSPQQQQQRAVYQAVVAVKLDCHNYVVAHAKPGVW  
PDLAYDSA KVMAAGLLKLGLFQNGTVDEIVDAGALAVFYPHGL  
GHGMGIDCHEIAHRAKGWPRGTCRGKKPHHSFVRFGRTLEKG  
VVITNEPGCYFIRPSYNAAFADPEKSKYRKRKVCERLRKTVGGV  
RIEDDLLITEDGCKVLSNIPKEIHRAKDEIEAFMAKCESKL

# *Why use a regular expression?*

To find a pattern

MALDVANRPMPKPEMFAAHRAKTLAELRKRKLEGVVLIYGFP  
EPTRDRINKEPVFRQESCFYWLTGVNEADCAYFLDIETGKEILFY  
PDIPQAYIIWFGELATIDDIQQQQQGFEDVRLMPKIQETLAEYK  
LKKIHTLRKRKIJKGYVAVKDKNEFIDVVGELRQIKDDDEMVLIQ  
YACDVNSFAVRDTFKVHPKMWEHQVMILKHYVDYYCRCFAF  
STIVCSGENCSILHYHHNNKFIEDGELILIDTGCEYNCAADNRTI  
PANGKFSPQQQQQRAVYQAVVAVKLDCHNYVVAHAKPGVW  
PDLAYDSA KVMAAGLLKLGLFQNGTVDEIVDAGALAVFYPHGL  
GHGMGIDCHEIAHRAKGWPRGTCRGKKPHHSFVRFGRTLEKG  
VVITNEPGCYFIRPSYNAAFADPEKSKYRKRKVCERLRKTVGGV  
RIEDDLLITEDGCKVLSNIPKEIHRAKDEIEAFMAKCESKL

# *Where do patterns come from?*

A sequence logo visualization showing the conservation of amino acids at each position of six protein sequences. The sequences are:

- F GCV . . W K A Q L L N E Y . . . . .
- / G E V . . W R G S W Q G E N . . . . .
- F G K V . . Y R A F W I G D E . . . . .
- F G T V . . Y K G K W H G D . . . . .
- S G T V . . Y K G V L E D D R H . . . . .
- V A V K I F P I Q . D K Q S W Q N E Y . . . . .
- V A V K I F S S R . D E K S W F R E T . . . . .
- V A V K A A R R H D P D E D I S Q T I E N V R . . . . .
- V A V K I L K V V . D P T P E Q F Q A F R N . . . . .
- V A V K K L E N V . R Q G K E V F Q A . . . . .
- E V Y S L P G M . K
- E L Y N T V M L . R
- Q E A K L F A M L . K
- E V A V L R K T . R
- E L S V I G R I . N

The logo uses a color scheme where each column represents an amino acid: F (blue), G (green), C (yellow), V (orange), S (red), P (purple), I (light blue), T (pink), A (light green), K (dark blue), R (dark red), D (dark purple), N (dark pink), M (brown), L (tan), H (light orange), E (light pink), W (light brown), Q (light purple), G (light green), Y (light orange), and M (light brown).

# *Where do patterns come from?*



# *Why use a regular expression?*

To find a pattern

MALDVANRPMPKPEMFAAHRAKTLAELRKRKLEGVVLIYGFPE  
PTRDRINKEPVFRQESCFYWLTGVNEADCAYFLDIETGKEILFYP  
DIPQAYIIWFGELATIDDIQQQQQGFEDVRLMPKIQETLAEYKL  
KKIHTLRKRKILKGY**VAVK**DKNEFIDVVGELRQIKDDDEMVLIQY  
ACDVNSFAVRDTFKVHPKMWEHQVMILKHYVDYYCRCFAFS  
TIVCSGENCSILHYHHNNKFIEDGELILIDTGCEYNCAADNTRTIP  
ANGKFSPQQQQQRAVYQAV**VAVK**LDCHNYVVAHAKPGVWP  
DLAYDSAKVMAAGLLKLGLFQNGTVDEIVDAGALAVFYPHGLG  
HGMGIDCHEIAHRAKGWPRGTCRGKKPHHSFVRFGRTLEKGV  
VITNEPGCYFIRPSYNAAFADPEKSKYRKRVCELRKTVGGVRI  
EDDLLITEDGCKVLSNIPKEIHRAKDEIEAFMAKKESKL

# *Why use a regular expression?*

MALDVANRPMPKPEMFAAHRAKTLAELRKRKLEGVVLIYGFPE  
PTRDRINKEPVFRQESC FYWLTGVNEADCAYFLDIETGKEILFYP  
DIPQAYIIWFGE LATIDDIQQQQQQGFEDVRLMPKIQETLAEYKL  
KKIHTLRKRKILKG YVAVKDKNEFIDVVGELRQIKDDDEMVLIQY  
ACDVNSFAVRDTFKVHPKMWEHQVMILKHYVDYYCRCFAFS  
TIVCSGENCSILHYHHNNKFIEDGELILIDTGCEYNCAADNTRTIP  
ANGKFSPQQQQQRAVYQAVVAVKLDCHNYVVAHAKPGVWP  
DLAYDSA KVMAAGLLKLGLFQNGTVDEIVDAGALAVFYPHGLG  
HGMGIDC HEIAHRAKGWPRGTCRGKKPHHSFVRFGRTLEKGV  
VITNEPGCYFIRPSYNAAFADPEKS YRKRKVCERLRKT VGGVRI  
EDDLLITEDGCKVLSNIPKEIHRAKDEIEAFMAKKDEL

**KDEL\$**

# *Why use a regular expression?*

MALDVANRPMPKPEMFAAHRAKTLAELRKRKLEGVVLIYGFPE  
PTRDRINKEPVFRQESC FYWLTGVNEADCAYFLDIETGKEILFYP  
DIPQAYIIWFGE LATIDDIQQQQQQGFEDVRLMPKIQETLAEYKL  
KKIHTLRKRKILKG YVAVKDKNEFIDVVGELRQIKDDDEMVLIQY  
ACDVNSFAVRDTFKVHPKMWEHQVMILKHYVDYYCRCFAFS  
TIVCSGENCSILHYHHNNKFIEDGELILIDTGCEYNCAADNTRTIP  
ANGKFSPQQQQQRAVYQAVVAVKLDCHNYVVAHAKPGVWP  
DLAYDSA KVMAAGLLKLGLFQNGTVDEIVDAGALAVFYPHGLG  
HGMGIDC HEIAHRAKGWPRGTCRGKKPHHSFVRFGRTLEKGV  
VITNEPGCYFIRPSYNAAFADPEKS YRKRKVCERLRKT VGGVRI  
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]