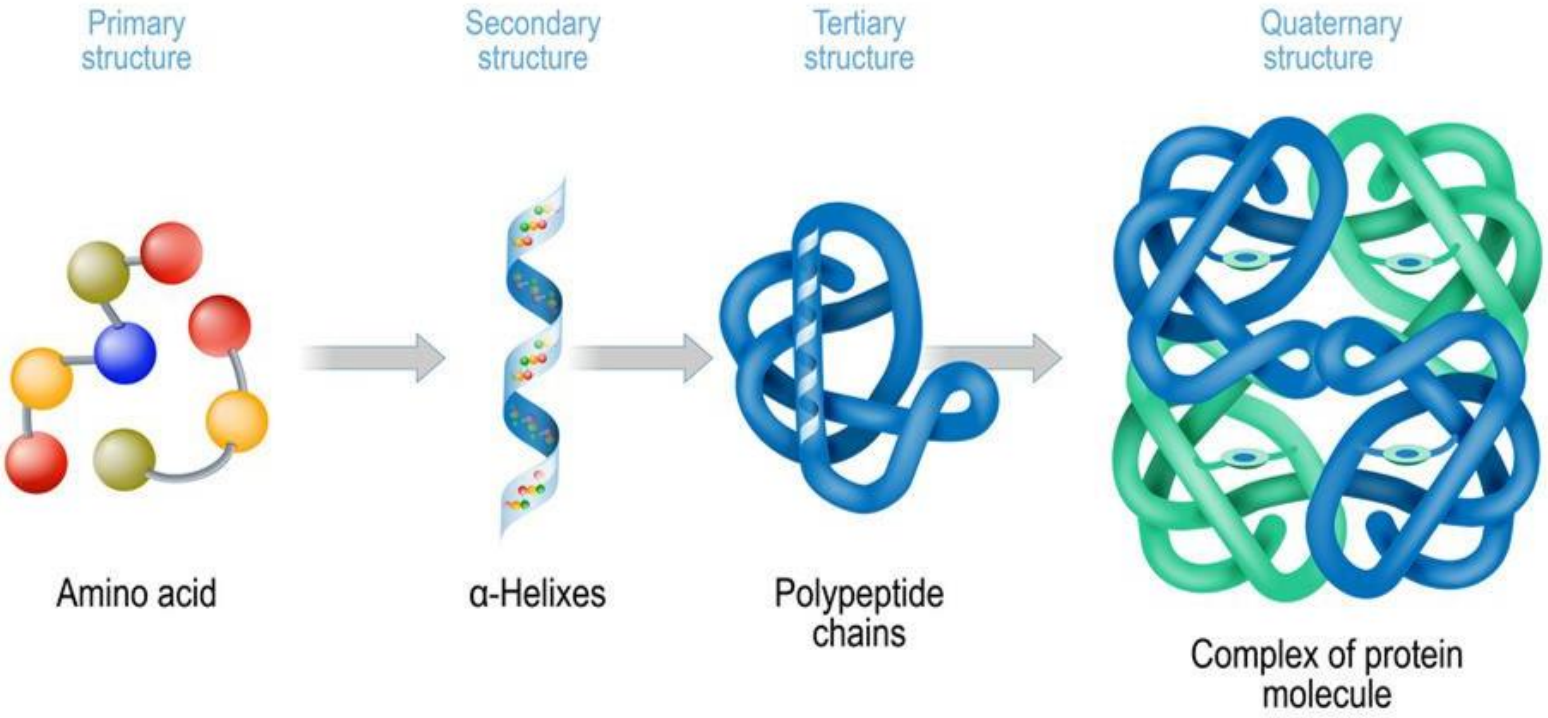
Patterns in nature

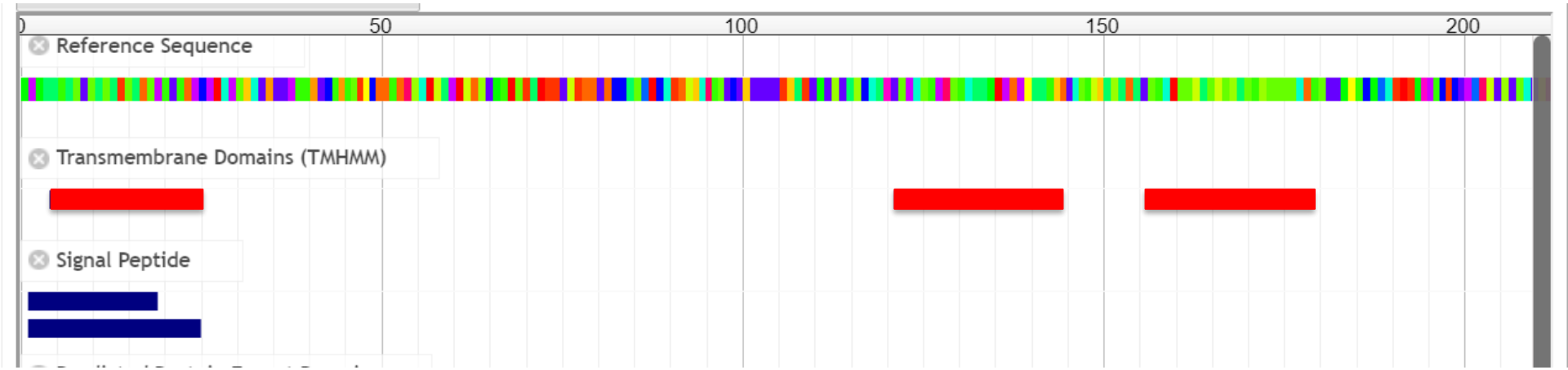
Predicted Protein Sequence

532 aa

```
MKLLNFPLSRYDVLITRFFLFASFLMVLFMCENFLKKSYMTSINIRNNKLCSFIHKNFENLEKYLTCDFVAYDLGMNDWK
SLKKKIIKNGEGEYNDMKDPSYMEKKHFEIEKSSHKIKSRQNEMNEQKSTLMNRPRKSKYNNTNKNNKVKGNLNKKKKKK
KNKKKAAQKYYTNKMGYSIGTNNNNNNNNNSNNNSNNNSNNNSNNNNNNNNNINIDSRRDKYKEVDKYYHESFNYIVDTN
NPSVDEQVLKKALMVFKKLDINKNKYIDYIEFEKNVNILSRMNEINKNILTYLFDMFDIDKDKKLNYTEFMSLNSYDFNY
```



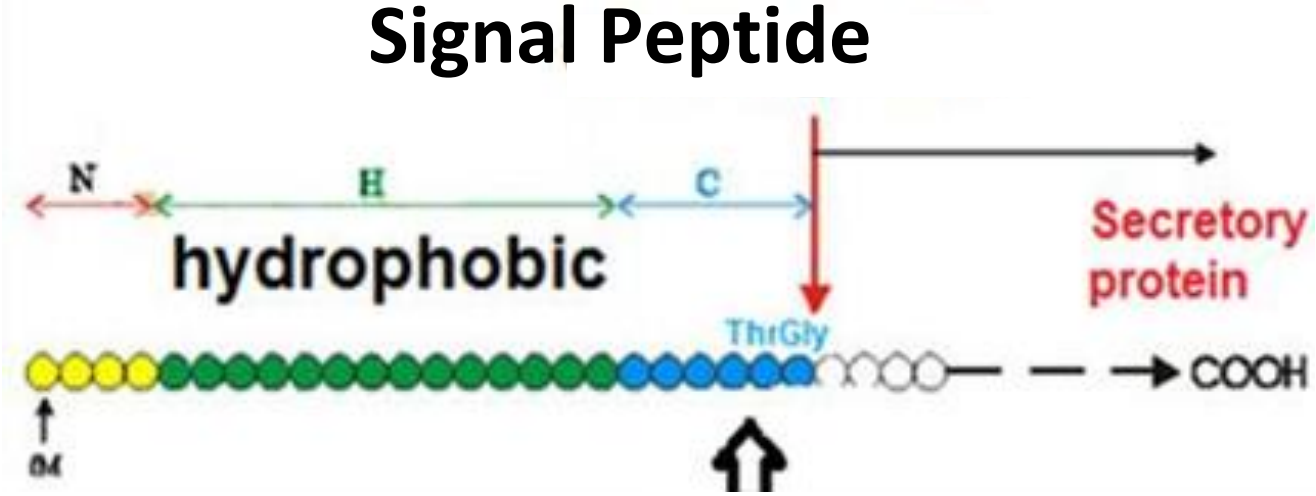
Patterns associated with function



Not exactly the same

```
NP_032136.1      MALPSNFLLGVCCFAWLCFLSSLSSQAS
NP_067704.1      MAFPSRFLLGVCCFAWLCLLISLSSQAS
ABW96807.1       MALPNKFLLWFYCFAWLCFPVSLGSQAS
XP_001504477.1   MALPSKFFLWFCCSAWLCFPISLGSQAS
ACK28140.1       MALPSKFLLWFCCLACMCFSASFGSQPP
NP_001001909.1   MALPRKFFLCFCCFALFCFPVSCGSQAS
                 ** *  * *   * *  *   *  **
Consensus        MALPNKFLLWFCCFAWLCFPISLGSQAS
```

Signal Peptide



InterPro Classification of protein families

Functional Characterization of Proteins

- classify proteins into families
- predicting domains and important sites
- predictive models, (signatures)
- several different databases that are members of the InterPro consortium.

http://www.ebi.ac.uk/interpro/

CATH-Gene3D

4.2.0

6k entries

CDD

3.17

15k entries

HAMAP

2020_01

2k entries

PANTHER

14.1

123k entries

Pfam

32.0

18k entries

PIRSF

3.10

3k entries

PRINTS

42.0

2k entries

PROSITE profiles

2019_11

1k entries

TIGRFAMs

15.0

4k entries

PROSITE patterns

2019_11

1k entries

SFLD

4

303 entries

SMART

7.1

1k entries

SUPERFAMILY

1.75

2k entries

Motifs
DNA and Protein

a nucleotide or amino-acid sequence pattern that is widespread and can have a biological significance.

Domains
Protein

a conserved part of a protein sequence and structure that can evolve, function, and exist independently of the rest of the protein chain.

- **Binding sites**
- **Enzyme activity**
- **Regulatory regions**

Domains at VEuPathDB

As we integrate data, we run programs that match or predict domains. We display this information on gene pages and create genome-wide searches of the program results

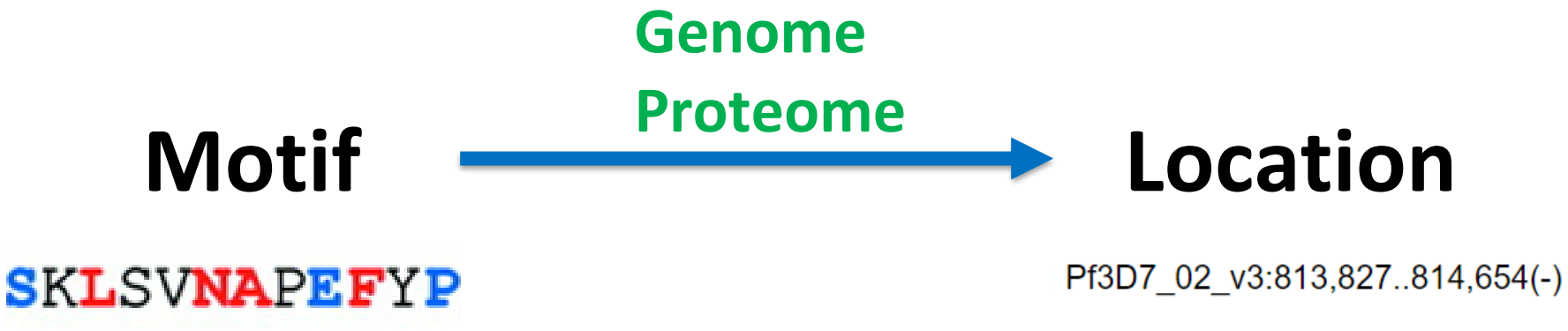
InterProScan - matches proteins against the InterPro protein signature databases

Signal P - predicts Signal Peptides in proteins

TMMHMM - predicts Transmembrane domains in proteins

How do we search for a motif in the VEuPathDB sea of DNA and protein?

Motif searches (text strings)



Regular expression is like another language

- a sequence of symbols and characters expressing a string or pattern to be searched for within a longer piece of text.
- Build in the ambiguity of a consensus sequence.
- Normal characters and symbols
  - Alphanumeric abc…ABC…0123...
  - Symbols punctuation to account for ambiguity -_ ,.;:=()/+ *%&{}[]?!$’^|\<>"@#
- 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

MALDVANRPMPKPEMFAAHRAKTLAELRKRKLEGVVLIYGFP
EPTRAHCDFEPVFRQESCFYWLTGVNEADCAYFLDIETGKEILF
YPDIPQAYIIWFGELATIDDIQQQQQGFEDVRLMPKIQETLAE
YKLKKIHTLPETCILKGYVAVKDKNEFIDVVGELRQIKDDDEMV
LIQYACDVNSFAVRDTFKKVHPKMWEHQVEANLIKHYVDYYC
RCFAFSTIVCSGENCSILHYHHNNKFIEDGELILIDTGCEYNCAA
DNTRTIPANGKFSPQQQQQRAVYQAVVAVKLDCHNYVVAH
AKPGVWPDLAYDSAKVMAAGLLKLGLFQNGTVDEIVDAGAL
AVFYPHGLGHGMGIDCHEIAHRAKGWPRGTCRGKKPHHSFV
RFGRTLEKGVVITNEPGCYFIRPSYNAAFADPEKSKYINKEVCER
LRKTVGGVRIEDDLLITEDGCKVLSNIPKEIHRAKDEIEAFMAKK
ESKL

Why use a regular expression?

To find a pattern

MALDVANRPMPKPEMFAAHRAKTLAELRKRKLEGVVLIYGFP
EPTRDRINKFEPVFRQESCFYWLTGVNEADCAYFLDIETGKEILF
YPDIPQAYIIWFGELATIDDI**QQQQQ**GFEDVRLMPKIQETLAE
YKLKKIHTLPETCILKGYVAVKDKNEFIDVVGELRQIKDDDEMV
LIQYACDVNSFAVRDTFKKVHPKMWEHQVMILKHYVDYYCR
CFAFSTIVCSGENCSILHYHHNNKFIEDGELILIDTGCEYNCAAD
NTRTIPANGKFSP**QQQQQ**RAVYQAVVAVKLDCHNYVVAHAK
PGVWPDLAYDSAKVMAAGLLKLGLFQNGTVDEIVDAGALAV
FYPHGLGHGMGIDCHEIAHRAKGWPRGTCRGKKPHHSFVRF
GRTLEKGVVITNEPGCYFIRPSYNAAFADPEKSKYINKEVCERLR
KTVGGVRIEDDLLITEDGCKVLSNIPKEIHRAKDEIEAFMAKKES
KL

Why use a regular expression?

To find a pattern

MALDVANRPMPKPEMFAAHRAKTLAEL**RKRK**LEGVVLIYGFP
EPTRDRINKEPVFRQESCFYWLTGVNEADCAYFLDIETGKEILF
YPDIPQAYIIWFGELATIDDI**QQQQQ**GFEDVRLMPKIQETLAE
YKLKKIHTL**RKRK**ILKGYVAVKDKNEFIDVVGELRQIKDDDEMV
LIQYACDVNSFAVRDTFKKVHPKMWEHQVMILKHYVDYYCR
CFAFSTIVCSGENCSILHYHHNNKFIEDGELILIDTGCEYNCAAD
NTRTIPANGKFSP**QQQQQ**RAVYQAVVAVKLDCHNYVVAHAK
PGVWPDLAYDSAKVMAAGLLKLGLFQNGTVDEIVDAGALAV
FYPHGLGHGMGIDCHEIAHRAKGWPRGTCRGKKPHHSFVRF
GRTLEKGVVITNEPGCYFIRPSYNAAFADPEKSKY**RKRK**VCERL
RKTVGGVRIEDDLLITEDGCKVLSNIPKEIHRAKDEIEAFMAKKE
SKL

Why use a regular expression?

To find a pattern

MALDVANRPMPKPEMFAAHRAKTLAEL**RKRK**LEGVVLIYGFP
EPTR**DRINK**EPVFRQESCFYWLTGVNEADCAYFLDIETGKEILF
YPDIPQAYIIWFGELATIDDI**QQQQQ**GFEDVRLMPKIQETLAE
YKLKKIHTL**RKRK**ILKGYVAVKDKNEFIDVVGELRQIKDDDEMV
LIQYACDVNSFAVRDTFKKVHPKMWEHQV**MILK**HYVDYYCR
CFAFSTIVCSGENCSILHYHHNNKFIEDGELILIDTGCEYNCAAD
NTRTIPANGKFSP**QQQQQ**RAVYQAVVAVKLDCHNYVVAHAK
PGVWPDLAYDSAKVMAAGLLKLGLFQNGTVDEIVDAGALAV
FYPHGLGHGMGIDCHEIAHRAKGWPRGTCRGKKPHHSFVRF
GRTLEKGVVITNEPGCYFIRPSYNAAFADPEKSKY**RKRK**VCERL
RKTVGGVRIEDDLLITEDGCKVLSNIPKEIHRAKDEIEAFMAKKE
SKL


VAVK

Why use a regular expression?

To find a pattern

MALDVANRPMPKPEMFAAHRAKTLAELRKRKLEGVVLIYGFP
EPTRDRINKEPVFRQESCFYWLTGVNEADCAYFLDIETGKEILF
YPDIPQAYIIWFGELATIDDIQQQQQGFEDVRLMPKIQETLAE
YKLKKIHTLRKRKILKGY**VAVK**DKNEFIDVVGELRQIKDDDEMV
LIQYACDVNSFAVRDTFKKVHPKMWEHQVMILKHYVDYYCR
CFAFSTIVCSGENCSILHYHHNNKFIEDGELILIDTGCEYNCAAD
NTRTIPANGKFSPQQQQQRAVYQAV**VAVK**LDCHNYVVAHA
KPGVWPDLAYDSAKVMAAGLLKLGLFQNGTVDEIVDAGALA
VFYPHGLGHGMGIDCHEIAHRAKGWPRGTCRGKKPHHSFVR
FGRTLEKGVVITNEPGCYFIRPSYNAAFADPEKSKYRKRKVCER
LRKTVGGVRIEDDLLITEDGCKVLSNIPKEIHRAKDEIEAFMAKK
ESKL

- **MLSTD**NVANRPMPKPEMF….

- 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]

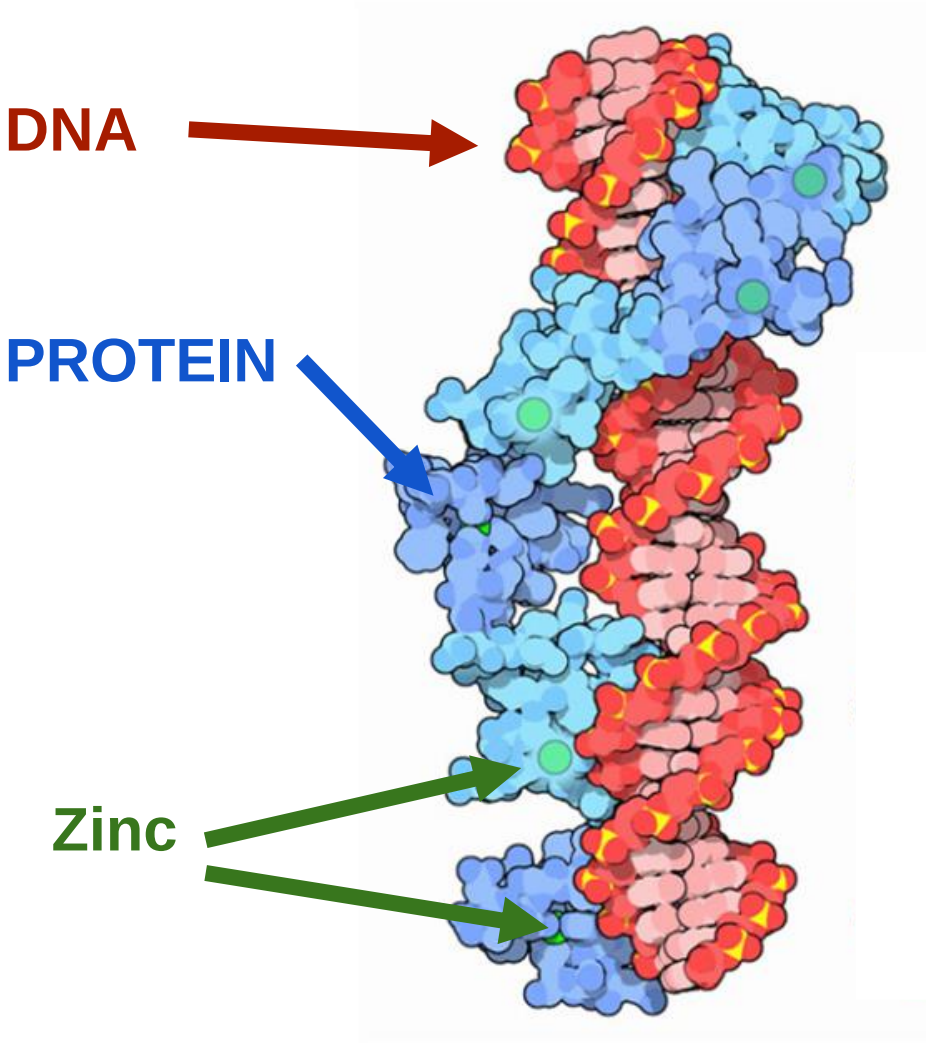
Useful RegEx help

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

Examples –
EcoR1 = GAATTC

AvaII = GGACC or GGTCC = GG[AT]CC

Zinc finger - zinc-containing domains found in a number of transcription factors



The zinc finger binding protein, transcription factor TFIIIA, binding to DNA

PDB101
https://pdb101.rcsb.org/motm/87

TFIIIA is a GATA-binding zinc finger protein

- DNA binding motif in the regulatory region of genes -
  - (A/T)GATA(A/G)
  - **[AT]GATA[AG]**
- GATA-type zinc finger domain -
  - C-x-[DNEHQSTI]-C-x(4,6)-[ST]-x(2)-[WM]-[HR]-[RKENAMSLPGQT]-x(3,4)-[GNEP]-x(3,6)-C-[NES]-[ASNR]-C
  - https://prosite.expasy.org/PS00344
  - **C.[DNEHQSTI]C.{4,6}[ST].{2}[WM][HR][RKENAMSLPGQT].{3,4}[GNEP].{3,6}C[NES][ASNR]C**