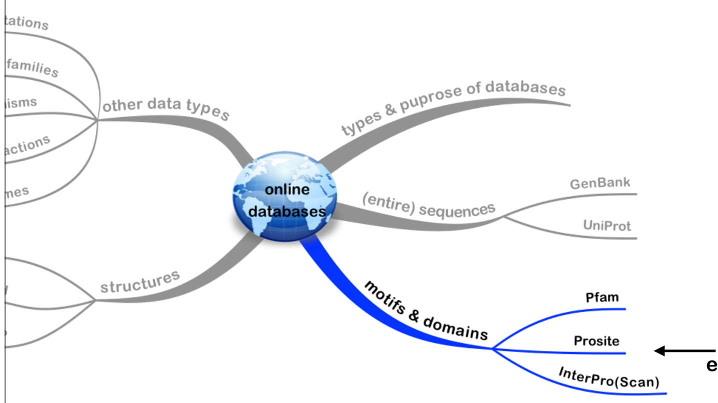
Bioinformatik
Stefanie Hartmann
Wintersemester 2019 / 2020, Universität Potsdam

Online resources (II), regular expressions
Nov 08, 2019

Overview



regular expressions

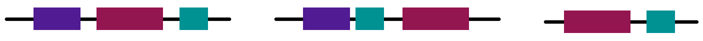
Protein domains & architectures

protein domains

- functional and evolutionary units of proteins
- have specific and conserved functions
- function independently or with other domains
- together determine the protein's function
- most eukaryotic proteins contain ≥ 2 domains

architecture

- ordered domain arrangement in a given protein



Motif and domain databases

sequence

cDNA, gDNA, protein

NCBI

```
MYALKRELWCVLLLCGAICTSPSQETHRRLRRGVRSYRVTCRDEKTQMIYQQHQSWLRPLLRGNRVEHCWCNDGQTQ
CHSVPVKSCSEPRCFNGGTCLQAIYFSDFVCQCPVGFIGRQCEIDARATCYEDQGITYRGTWSTTESGAECVNWNTS
GLASMPYNGRRPDAVKLGLGNHNYCRNPDKDSKPWCYIFKAEKYSPDFCSTPACTKEKEECYTGKGLDYRGTRSLTM
SGAFCLPWNSLVLMGKIYTAWNSNAQTLGLGKHNYCRNPDGDTQPWCHVLKDHKLTWEYCDLPQCVTCGLRQYKEPQ
FRIKGGLYADITSHPWQAAIFVKNRRSPGERFLCGGILISSCWVLSAAHCFQERFPPHHVRVVLGRTYRLVPGEEEQ
AFEVEKYIVHKEFDDDTYDNDIALLQLKSDSLTCAQESDAVRTVCLPEANLQLPDWTECELSGYGKHEASSPFYSER
LKEAHVRLYPSSRCTSKHLFNKTITNNMLCAGDTRSGGDNANLHDACQGDSGGPLVCMKGNHMTLVGVISWGLGCGQ
KDVPGVYTKVTNYLNWIRDNTRP
```

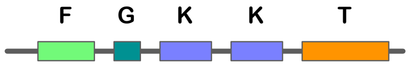
function

UniProt

convert zymogen plasminogen to plasmin

motifs/domains

- 1 fibronectin
- 1 growth factor
- 2 Kringle
- 1 trypsin

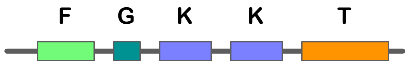


Motif and domain databases

- purposes:
  - group sequences that contain conserved patterns into “families”
  - annotate patterns and their families
  - scan new sequences against stored patterns
- identification/assignment of patterns:
  - Hidden Markov models (e.g., Pfam) (statistical models of conserved sequences)
  - regular expressions (e.g., Prosite) (syntax to describe patterns)
  - and more

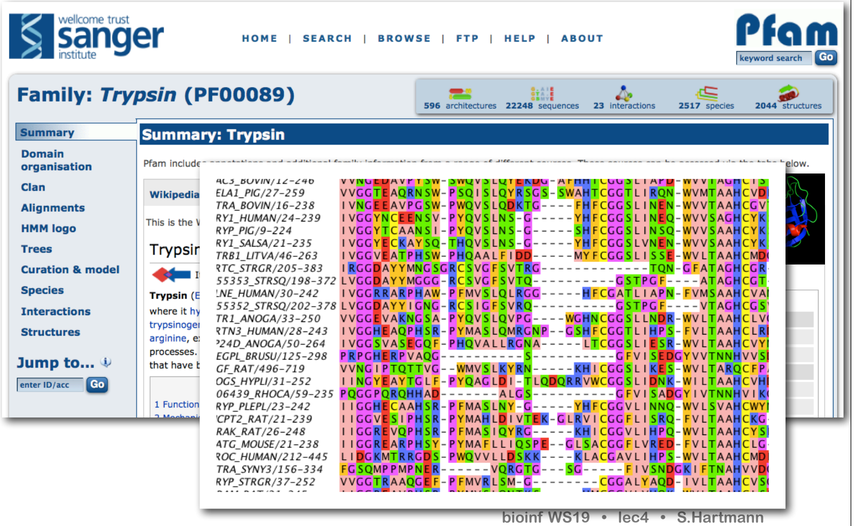
Pfam

- entries are based on functional domains



- available information:
  - domain combinations containing a given domain
  - sequences in which the domain occurs
  - species in which the domain has been found
  - 3D structures, if available
  - …

A Pfam entry



Family: *Trypsin* (PF00089)

Summary: Trypsin

Pfam include[...] below.

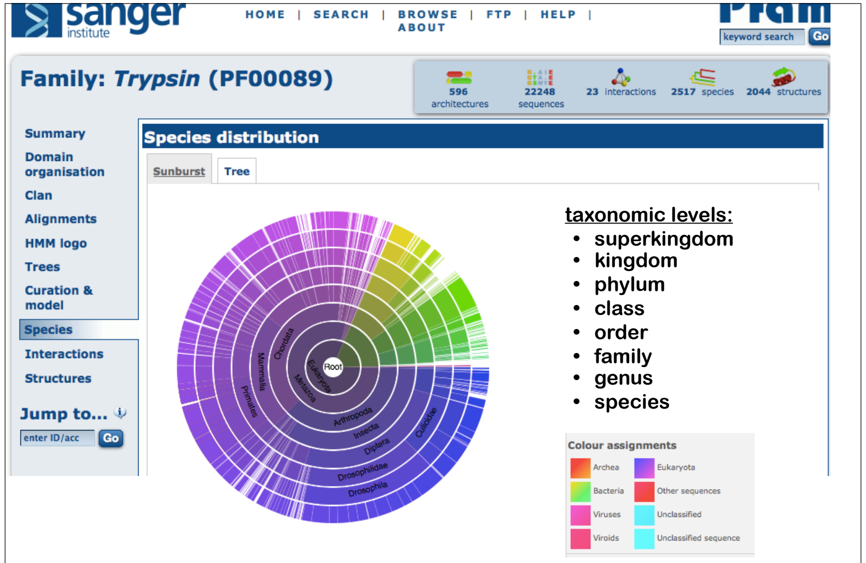
This is the V

Trypsir

Trypsin (E
where it hy
trypsinoger
arginine, e
processes.
that have b

1 Function

```
AC3_BOVIN/12-246      VVNGEDAVPYSW-SWQVSLQYEKDG-AFHHTCGGSLIAPD-WVVTAGHCIS
ELA1_PIG/27-259       VVGGTEAQRNSW-PSQISLQYRSGS-SWAHTCGGTLIRQN-WVMTAAHCVD
TRA_BOVIN/16-238      IVNGEEAVPGSW-PWQVSLQDKTG----FHFCGGSLINEN-WVVTAAHCGV
RY1_HUMAN/24-239      IVGGYNCEENSV-PYQVSLNS-G-----YHFCGGSLINEQ-WVVSAGHCYK
RYP_PIG/9-224         IVGGYTCAANSI-PYQVSLNS-G-----SHFCGGSLINSQ-WVVSAAHCYK
RY1_SALSA/21-235      IVGGYECKAYSQ-THQVSLNS-G-----YHFCGGSLVNEN-WVVSAAHCYK
TRB1_LITVA/46-263     IVGGVEATPHSW-PHQAALFIDD-----MYFCGGSLISSE-WVLTAAHCMD
RTC_STRGR/205-383     IRGGDAYYMNGSGRCSVGFSVTRG-------------TQN-GFATAGHCGR
55353_STRSQ/198-372   LVGGDAYYMGGG-RCSVGFSVTQ-----------GSTPGF----ATAGHCGT
LNE_HUMAN/30-242      IVGGRRARPHAW-PFMVSLQLRGG-----HFCGATLIAPN-FVMSAAHCVA
55352_STRSQ/202-378   LVGGDAYYIGNG-RCSIGFSVRQ-----------GSTPGF----VTAGHCGS
TR1_ANOGA/33-250      VVGGEVAKNGSA-PYQVSLQVPG----WGHNCGGSLLNDR-WVLTAAHCLV
RTN3_HUMAN/28-243     IVGGHEAQPHSR-PYMASLQMRGNP--GSHFCGGTLIHPS-FVLTAAHCLR
P24D_ANOGA/50-264     IVGGSVASEGQF-PHQVALLRGNA-----LTCGGSLIESR-WVLTAAHCVY
EGPL_BRUSU/125-298    PRPGHERPVAQG----------S-----------GFVISEDGYVVTNNHVVS
GF_RAT/496-719        VVNGIPTQTTVG--WMVSLKYRN-----KHICGGSLIKES-WVLTARQCFP
OGS_HYPLI/31-252      IINGYEAYTGLF-PYQAGLDI-TLQDQRRVWCGGSLIDNK-WILTAAHCVH
06439_RHOCA/59-235    PQGGPQRQHHAD-------ALGS-----------GFVISADGYIVTNNHVIK
RYP_PLEPL/23-242      IIGGHECAAHSR-PFMASLNY-G-----YHFCGGVLINNQ-WVLSVAHCWY
CPT2_RAT/21-239       IIGGVESIPHSR-PYMAHLDIVTEK-GLRVICGGFLISRQ-FVLTAAHCKG
RAK_RAT/26-248        IIGGREVQPHSR-PFMASIQYRG-----KHICGGVLIHPQ-WVLTAAHCYS
ATG_MOUSE/21-238      IIGGREARPHSY-PYMAFLLIQSPE--GLSACGGFLVRED-FVLTAAHCLG
ROC_HUMAN/212-445     LIDGKMTRRGDS-PWQVVLLDSKK----KLACGAVLIHPS-WVLTAAHCMD
TRA_SYNY3/156-334     FGSQMPPMPNER------VQRGTG---SG------FIVSNDGKIFTNAHVVD
RYP_STRGR/37-252      VVGGTRAAQGEF-PFMVRLSM-G---------CGGALYAQD-IVLTAAHCVS
```



Family: *Trypsin* (PF00089)

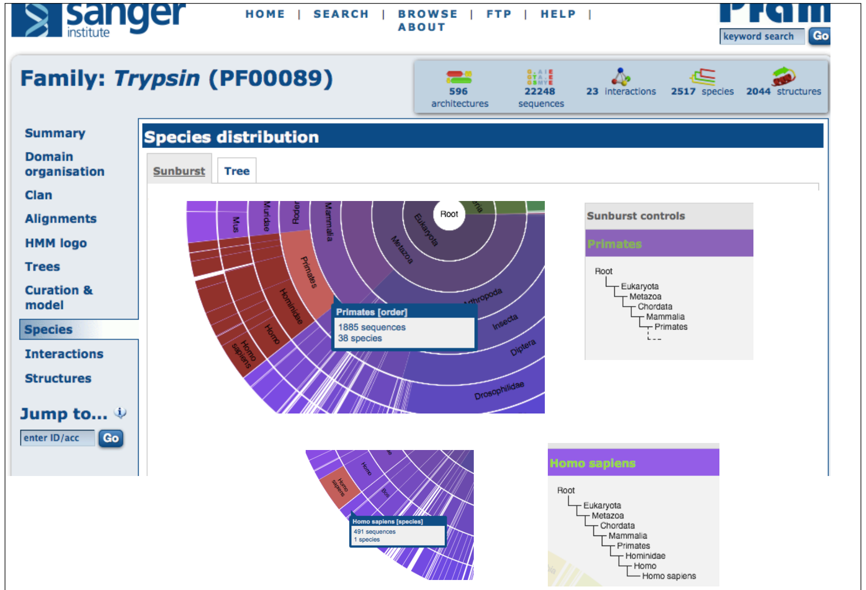
Species distribution

taxonomic levels:

- superkingdom
- kingdom
- phylum
- class
- order
- family
- genus
- species

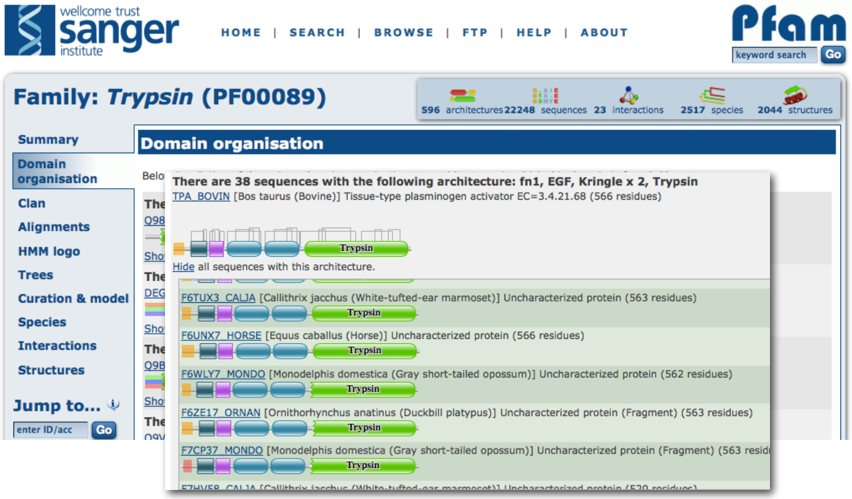
Colour assignments

🟥	Archea	🟪	Eukaryota
🟩	Bacteria	🟥	Other sequences
🟪	Viruses	🟦	Unclassified
🟥	Viroids	🟦	Unclassified sequence



Family: *Trypsin* (PF00089)

Species distribution



wellcome trust sanger institute
HOME | SEARCH | BROWSE | FTP | HELP | ABOUT
Pfam
keyword search
Go

Family: *Trypsin* (PF00089)

Domain organisation

There are 38 sequences with the following architecture: fn1, EGF, Kringle x 2, Trypsin

TPA_BOVIN [Bos taurus (Bovine)] Tissue-type plasminogen activator EC=3.4.21.68 (566 residues)

Hide all sequences with this architecture.

F6TUX3_CALJA [Callithrix jacchus (White-tufted-ear marmoset)] Uncharacterized protein (563 residues)

F6UNX7_HORSE [Equus caballus (Horse)] Uncharacterized protein (566 residues)

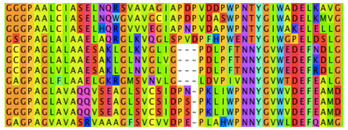
F6WLY7_MONDO [Monodelphis domestica (Gray short-tailed opossum)] Uncharacterized protein (562 residues)

F6ZE17_ORNAN [Ornithorhynchus anatinus (Duckbill platypus)] Uncharacterized protein (Fragment) (563 residues)

F7CP37_MONDO [Monodelphis domestica (Gray short-tailed opossum)] Uncharacterized protein (Fragment) (563 resid

Pfam

identify & align a representative set of sequences for the domain

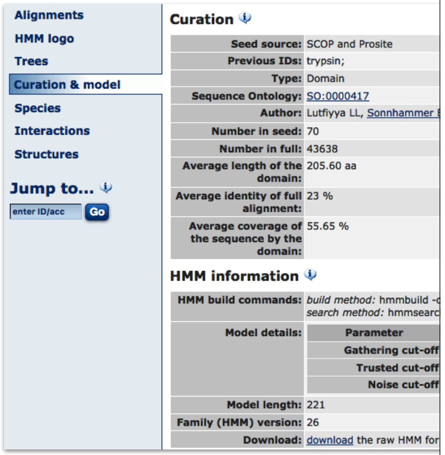


HMMer software

HMM: statistical model of sequences in the alignment

HMMer software

- efficiently identify additional sequences that belong to this family,
- add these to the alignment



Curation

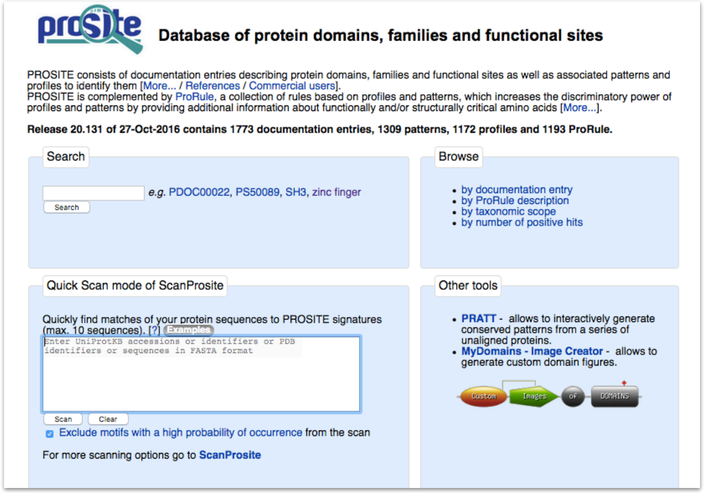
Seed source:	SCOP and Prosite
Previous IDs:	trypsin;
Type:	Domain
Sequence Ontology:	SO:0000417
Author:	Lutfiyya LL, Sonnhammer E
Number in seed:	70
Number in full:	43638
Average length of the domain:	205.60 aa
Average identity of full alignment:	23 %
Average coverage of the sequence by the domain:	55.65 %

HMM information

HMM build commands:	*build method:* hmmbuild -c *search method:* hmmsearc
Model details:	Parameter Gathering cut-off Trusted cut-off Noise cut-off
Model length:	221
Family (HMM) version:	26
Download:	download the raw HMM for

PROSITE

- http://www.expasy.ch/prosite/
- available at the Swiss Institute of Bioinformatics
- a database of biologically significant sites and patterns:
  - group sequences that contain conserved patterns into families
  - scan sequences against for allow to identify to which known family of protein (if any) a given sequence belongs
- based on regular expressions, a language for specifying patterns



Database of protein domains, families and functional sites

PROSITE consists of documentation entries describing protein domains, families and functional sites as well as associated patterns and profiles to identify them [More... / References / Commercial users].
PROSITE is complemented by ProRule, a collection of rules based on profiles and patterns, which increases the discriminatory power of profiles and patterns by providing additional information about functionally and/or structurally critical amino acids [More...].

Release 20.131 of 27-Oct-2016 contains 1773 documentation entries, 1309 patterns, 1172 profiles and 1193 ProRule.

Search

e.g. PDOC00022, PS50089, SH3, zinc finger

Search

Browse

- by documentation entry
- by ProRule description
- by taxonomic scope
- by number of positive hits

Quick Scan mode of ScanProsite

Quickly find matches of your protein sequences to PROSITE signatures (max. 10 sequences). [?] Examples

Enter UniProtKB accessions or identifiers or PDB identifiers or sequences in FASTA format

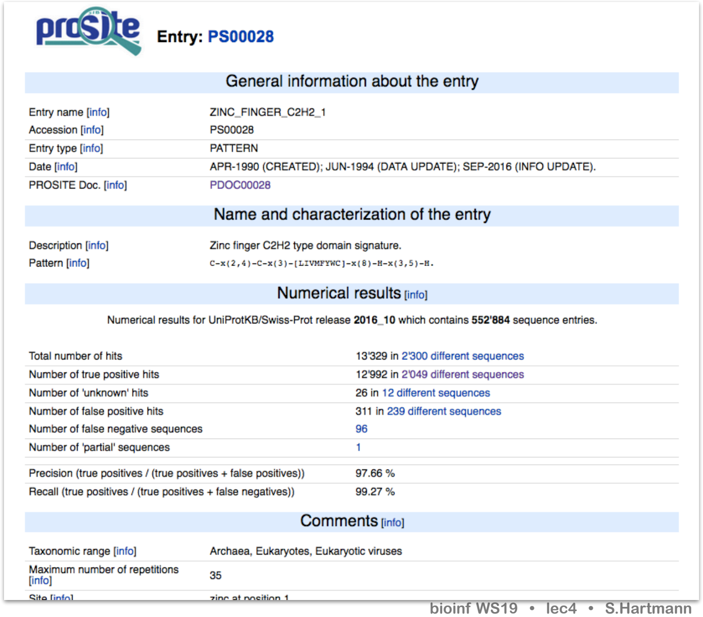
Scan Clear

☑ Exclude motifs with a high probability of occurrence from the scan

For more scanning options go to **ScanProsite**

Other tools

- **PRATT** - allows to interactively generate conserved patterns from a series of unaligned proteins.
- **MyDomains - Image Creator** - allows to generate custom domain figures.



Entry: PS00028

General information about the entry

Entry name [info]	ZINC_FINGER_C2H2_1
Accession [info]	PS00028
Entry type [info]	PATTERN
Date [info]	APR-1990 (CREATED); JUN-1994 (DATA UPDATE); SEP-2016 (INFO UPDATE).
PROSITE Doc. [info]	PDOC00028

Name and characterization of the entry

Description [info]	Zinc finger C2H2 type domain signature.
Pattern [info]	C-x(2,4)-C-x(3)-[LIVMFYWC]-x(8)-H-x(3,5)-H.

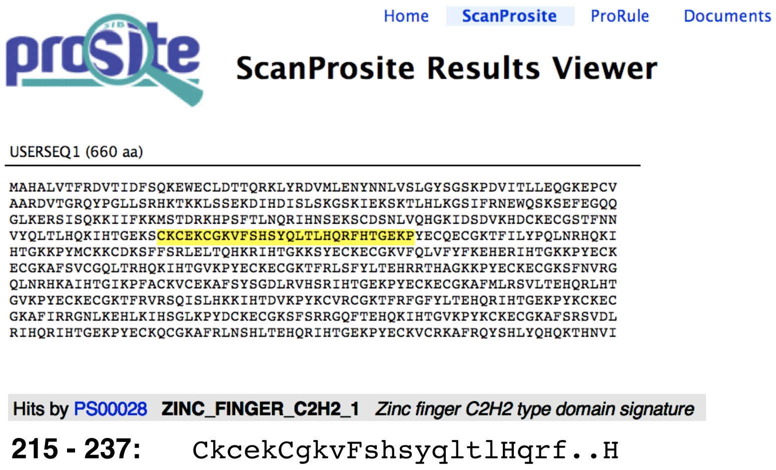
Numerical results [info]

Numerical results for UniProtKB/Swiss-Prot release **2016_10** which contains **552'884** sequence entries.

Total number of hits	13'329 in 2'300 different sequences
Number of true positive hits	12'992 in 2'049 different sequences
Number of 'unknown' hits	26 in 12 different sequences
Number of false positive hits	311 in 239 different sequences
Number of false negative sequences	96
Number of 'partial' sequences	1
Precision (true positives / (true positives + false positives))	97.66 %
Recall (true positives / (true positives + false negatives))	99.27 %

Comments [info]

Taxonomic range [info]	Archaea, Eukaryotes, Eukaryotic viruses
Maximum number of repetitions [info]	35
Site [info]	zinc at position 1



Home ScanProsite ProRule Documents

ScanProsite Results Viewer

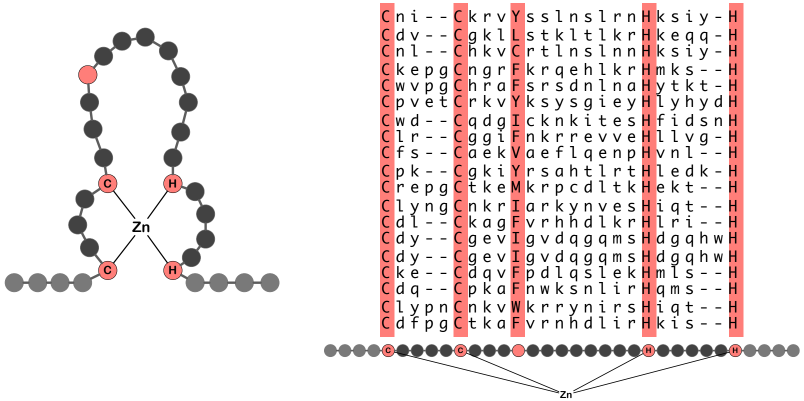
USERSEQ1 (660 aa)

```
MAHALVTFRDVTIDFSQKEWECLDTTQRKLYRDVMLENYNNLVSLGYSGSKPDVITLLEQGKEPCV
AARDVTGRQYPGLLSRHKTKKLSSEKDIHDISLSKGSKIEKSKTLHLKGSIFRNEWQSKSEFEGQQ
GLKERSISQKKIIFKKMSTDRKHPSFTLNQRIHNSEKSCDSNLVQHGKIDSDVKHDCKECGSTFNN
VYQLTLHQKIHTGEKSCKCEKCGKVFSHSYQLTLHQRFHTGEKPYECQECGKTFILYPQLNRHQKI
HTGKKPYMCKKCDKSFFSRLELTQHKRIHTGKKSYECKECGKVFQLVFYFKEHERIHTGKKPYECK
ECGKAFSVCGQLTRHQKIHTGVKPYECKECGKTFRLSFYLTEHRRTHAGKKPYECKECGKSFNVRG
QLNRHKAIHTGIKPFACKVCEKAFSYSGDLRVHSRIHTGEKPYECKECGKAFMLRSVLTEHQRLHT
GVKPYECKECGKTFRVRSQISLHKKIHTDVKPYKCVRCGKTFRFGFYLTEHQRIHTGEKPYKCKEC
GKAFIRRGNLKEHLKIHSGLKPYDCKECGKSFSRRGQFTEHQKIHTGVKPYKCKECGKAFSRSVDL
RIHQRIHTGEKPYECKQCGKAFRLNSHLTEHQRIHTGEKPYECKVCRKAFRQYSHLYQHQKTHNVI
```

Hits by PS00028 **ZINC_FINGER_C2H2_1** *Zinc finger C2H2 type domain signature*

215 - 237: CkcekCgkvFshsyqltlHqrf..H

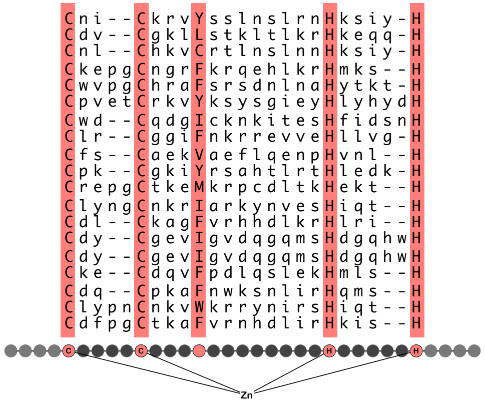
C2H2 zinc finger DNA binding protein



```
Cni--CkrvYsslnslrnHksiy-H
Cdv--CgklLstkltlkrHkeqq-H
Cnl--ChkvCrtlnslnnHksiy-H
CkepgCngrFkrqehlkrHmks--H
CwvpgChraFsrsdnlnaHytkt-H
CpvetCrkvYksysgieyHlyhydH
Cwd--CqdgIcknkitesHfidsnH
Clr--CggiFnkrrevveHllvg-H
Cfs--CaekVaeflqenpHvnl--H
Cpk--CgkiYrsahtlrtHledk-H
CrepgCtkeMkrpcdltkHekt--H
ClyngCnkrIarkynvesHiqt--H
Cdl--CkagFvrhhdlkrHlri--H
Cdy--CgevIgvdqgqmsHdgqhwH
Cdy--CgevIgvdqgqmsHdgqhwH
Cke--CdqvFpdlqslekHmls--H
Cdq--CpkaFnwksnlirHqms--H
ClypnCnkvWkrrynirsHiqt--H
CdfpgCtkaFvrnhdlirHkis--H
```

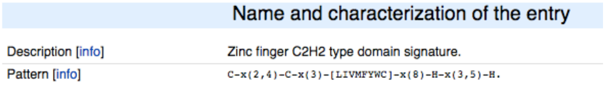
C2H2 zinc finger DNA binding protein

C followed by
any 2 to 4 letters followed by
C followed by
any 3 letters followed by
C followed by
one of L,I,V,M,F,Y,W,C followed by
any 8 letters followed by
H followed by
any 3 to 5 letters followed by
H



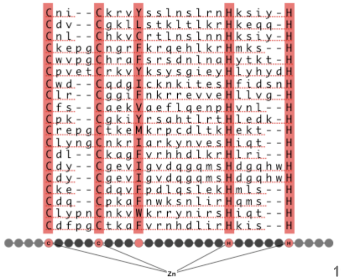
```
Cni--CkrvYsslnslrnHksiy-H
Cdv--CgklLstkltlkrHkeqq-H
Cnl--ChkvCrtlnslnnHksiy-H
CkepgCngrFkrqehlkrHmks--H
CwvpgChraFsrsdnlnaHytkt-H
CpvetCrkvYksysgieyHlyhydH
Cwd--CqdgIcknkitesHfidsnH
Clr--CggiFnkrrevveHllvg-H
Cfs--CaekVaeflqenpHvnl--H
Cpk--CgkiYrsahtlrtHledk-H
CrepgCtkeMkrpcdltkHekt--H
ClyngCnkrIarkynvesHiqt--H
Cdl--CkagFvrhhdlkrHlri--H
Cdy--CgevIgvdqgqmsHdgqhwH
Cdy--CgevIgvdqgqmsHdgqhwH
Cke--CdqvFpdlqslekHmls--H
Cdq--CpkaFnwksnlirHqms--H
ClypnCnkvWkrrynirsHiqt--H
CdfpgCtkaFvrnhdlirHkis--H
```

A Prosite regular expression



Name and characterization of the entry	
Description [info]	Zinc finger C2H2 type domain signature.
Pattern [info]	C-x(2,4)-C-x(3)-[LIVMFYWC]-x(8)-H-x(3,5)-H.

C-x(2,4)-C-x(3)-[LIVMFYWC]-x(8)-H-x(3,5)-H



Regular expressions

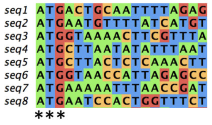
a powerful language for specifying patterns

used in different contexts and in different "dialects"

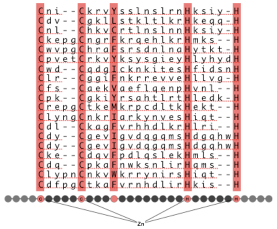
- linux terminals
- text editors
- protein motif databases (e.g., Prosite)
- ...

patterns

- exact and inexact patterns



```
seq1 ATGACTGCAATTTTAGAG
seq2 ATGAATGTTTTATCATGT
seq3 ATGGTAAAACTTCGTTTA
seq4 ATGCTTAATATATTTAAT
seq5 ATGCTTACTCTCAAACTT
seq6 ATGGTAACCATTAGAGCC
seq7 ATGAAAAATTTAACCGAT
seq8 ATGAATCCACTGGTTTCT
     ***
```



Regular expressions

a powerful language for specifying patterns

used in different contexts and in different "dialects"

- linux terminals
- text editors
- protein motif databases (e.g., Prosite)
- ...

patterns

- exact and inexact patterns
- in biological sequences: proteins, DNA
- in everyday life

Linux regular expressions, using grep

```
grep  Mus  species.txt
```

Mus musculus

```
grep  Sus  species.txt
```

Sus scrofa

regular expressions (regex)

Linux regular expressions

characters

- exact matches: ATG A Mus Sus 2 37

metacharacters

- characters with special meaning: [ ] . ? * + {} \ () | ^ $

metasymbols

- sequences of characters with special meaning
  \t \n \w

Linux regular expressions, using grep

```
grep      [MS]us        species.txt

grep -E 'MT{2,}NG'     proteins.fasta
```

regular expressions (regex)

grep parameters

-E allows “extended” regex syntax

-o -i -c …

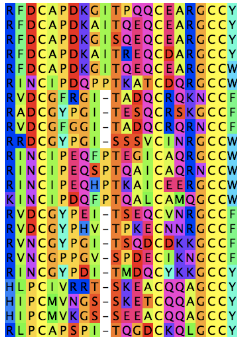
quotes to protect special characters

Translation: Prosite to grep/linux

Trefoil motif:

plays a role in the renewal and pathology of mucous epithelia

[KRH]-x(2)-C-x-[FYPSTV]-x(3,4)-[ST]-x(3)-C-x(4)-C-C-[FYWH]



```
RFDCAPDKGITPQQCEARGCCY
RFDCAPDKAITQEQCEARGCCY
RFDCAPDKGISQEQCEARGCCY
RFDCAPDKAITREQCDARGCCY
RFDCAPDKGITQEQCEARGCCW
RINCIPDQPPTKATCDQRGCCW
RVDCGFRGI-TADQCRQKNCCF
RADCGYPGI-TESQCRSKGCCF
RVDCGFGGI-TADQCRQRNCCF
RRDCGYPGI-SSSVCINRGCCW
RINCIPEQFPTEGICAQRGCCW
RINCIPEQSPTQAICAQRNCCW
RINCIPEQHPTKAICEERGCCW
KINCIPDQFPTQALCAMQGCCW
RVDCGYPEI-TSEQCVNRGCCF
RVDCGYPHV-TPKECNNRGCCF
RVNCGYPGI-TSQDCDKKGCCF
RVNCGPPGV-SPDECIKNGCCF
RINCGYPDI-TMDQCYKKGCCY
HLPCIVRRT-SKEACQQAGCCY
HIPCMVNGS-SKETCQQAGCCY
HIPCMVKGS-SEEACQQAGCCY
RLPCAPSPI-TQGDCKQLGCCY
```

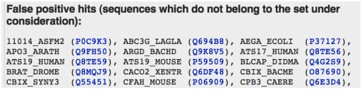
Prosite regex	grep regex
One of the amino acids K, R, or H	[KRH]
any two amino aicds	.{2}
a C	C
any one amino acid	.{1}
one of F, Y, P, S, T, or V	[FYPSTV]
any three or four amino acids	.{3,4}
either S or T	[ST]
any three amino acids	.{3}
a C	C
any four amino acids	.{4}
two Cs	C{2}
any one of the amino acids F, Y, W, H	[FYWH]

Translation: Prosite to grep/linux

C-x(2,4)-C-x(3)-[LIVMFYWC]-x(8)-H-x(3,5)-H

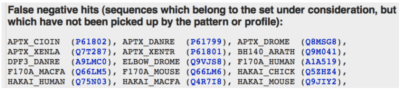
Some disadvantages of regex

- too rigid to pick up divergent sequences
- short patterns will find false positives



False positive hits (sequences which do not belong to the set under consideration):

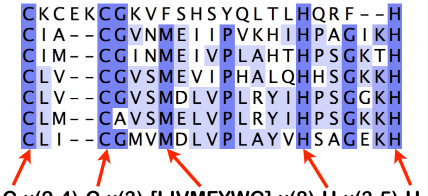
```
11014_ASFM2 (P0C9K3), ABC3G_LAGLA (Q694B8), AEGA_ECOLI  (P37127),
APO3_ARATH  (Q9FH50), ARGD_BACHD  (Q9K8V5), ATS17_HUMAN (Q8TE56),
ATS19_HUMAN (Q8TE59), ATS19_MOUSE (P59509), BLCAP_DIDMA (Q4G2S9),
BRAT_DROME  (Q8MQJ9), CACO2_XENTR (Q6DF48), CBIX_BACME  (O87690),
CBIX_SYNY3  (Q55451), CFAH_MOUSE  (P06909), CPB3_CAERE  (Q6E3D4),
```



False negative hits (sequences which belong to the set under consideration, but which have not been picked up by the pattern or profile):

```
APTX_CIOIN  (P61802), APTX_DANRE  (P61799), APTX_DROME  (Q8MSG8),
APTX_XENLA  (Q7T287), APTX_XENTR  (P61801), BH140_ARATH (Q9M041),
DPF3_DANRE  (A9LMC0), ELBOW_DROME (Q9VJS8), F170A_HUMAN (A1A519),
F170A_MACFA (Q66LM5), F170A_MOUSE (Q66LM6), HAKAI_CHICK (Q5ZHZ4),
HAKAI_HUMAN (Q75N03), HAKAI_MACFA (Q4R7I8), HAKAI_MOUSE (Q9JIY2),
```

- cannot include information about relative frequencies



```
CKCEKCGKVFSHSYQLTLHQRF--H
CIA--CGVNMEIIPVKHIHPAGIKH
CIM--CGINMEIVPLAHTHPSGKTH
CLV--CGVSMEVIPHALQHHSGKKH
CLV--CGVSMDLVPLRYIHPSGGKH
CLM--CAVSMELVPLRYIHPSGKKH
CLI--CGMVMDLVPLAYVHSAGEKH
```

C-x(2,4)-C-x(3)-[LIVMFYWC]-x(8)-H-x(3,5)-H

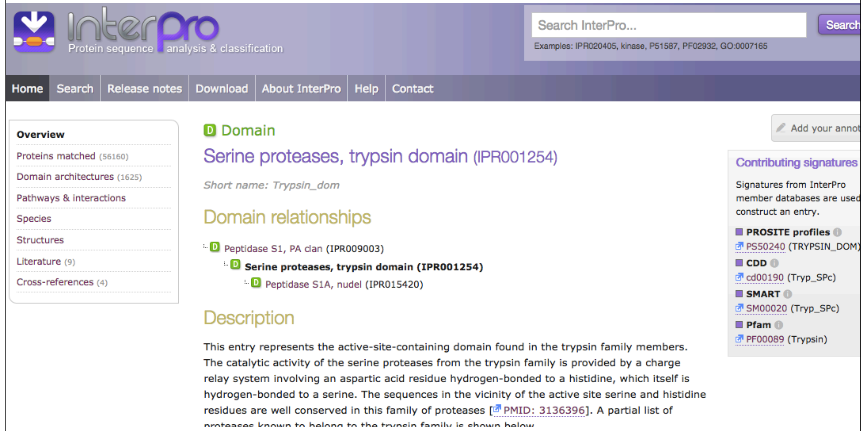
Motif and domain databases

Database	URL
PROSITE	http://www.expasy.org/prosite/
PRINTS	http://bioinf.man.ac.uk/dbbrowser/PRINTS/
Pfam	http://www.sanger.ac.uk/Software/Pfam/
ProDom	http://prodes.toulouse.inra.fr/prodom/current/html/home.php
BLOCKS	http://www.blocks.fhcrc.org/
SMART	http://smart.embl-heidelberg.de/
TIGRfam	http://www.tigr.org/TIGRFAMs/index.shtml
SUPERFAMILY	http://supfam.mrc-lmb.cam.ac.uk/SUPERFAMILY/
SBASE	http://www3.icgeb.trieste.it/~sbasesrv/main.html
CDD	http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml
...	...
...	

InterPro

- an integrated documentation resource for protein families, domains and sites.
- it combines a number of databases that use different approaches to classify domains
  - PROSITE (regular expressions and profiles)
  - Gene3D, PANTHER, PIRSF, Pfam, SMART, SUPERFAMILY and TIGRFAMs (hidden Markov models (HMMs))
  - PRINTS (groups of aligned, un-weighted motifs)
  - ProDom (PSI-BLAST approach)
- entries are
  - linked to entries in UniProt (and other databases)

InterPro



Overview

- Proteins matched (56160)
- Domain architectures (1625)
- Pathways & interactions
- Species
- Structures
- Literature (9)
- Cross-references (4)

D Domain

Serine proteases, trypsin domain (IPR001254)

Short name: Trypsin_dom

Domain relationships

- D Peptidase S1, PA clan (IPR009003)
  - D **Serine proteases, trypsin domain (IPR001254)**
    - D Peptidase S1A, nudel (IPR015420)

Description

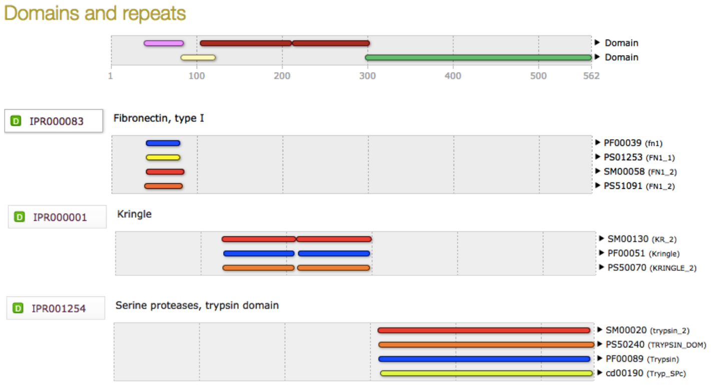
This entry represents the active-site-containing domain found in the trypsin family members. The catalytic activity of the serine proteases from the trypsin family is provided by a charge relay system involving an aspartic acid residue hydrogen-bonded to a histidine, which itself is hydrogen-bonded to a serine. The sequences in the vicinity of the active site serine and histidine residues are well conserved in this family of proteases [PMID: 3136396]. A partial list of proteases known to belong to the trypsin family is shown below.

Contributing signatures

Signatures from InterPro member databases are used construct an entry.

- **PROSITE profiles**
  - PS50240 (TRYPSIN_DOM)
- **CDD**
  - cd00190 (Tryp_SPc)
- **SMART**
  - SM00020 (Tryp_SPc)
- **Pfam**
  - PF00089 (Trypsin)

InterproScan



Today’s exercise

1. regular expressions
   - patterns in protein sequences
     - thionins
     - zinc fingers
   - patterns in DNA sequences
     - OPTIONAL: restriction enzyme recognition sites

proteins.fasta

dna.fasta

2. domain databases: Pfam

NOTE: open these two files in the terminal or with text-only editors!

Terms and concepts

- domain databases
  - organization
  - purpose
  - examples
- regular expressions
  - definition, application
  - relationship to protein domains
  - syntax Prosite vs. linux terminal
    - `grep (-o, -E)`
    - metacharacters `[] . ? * + {} \`