Basel, November 2018, christian.sigrist@sib.swiss ivo.pedruzzi@sib.swiss

PROSITE and HAMAP

Remember to carefully read the documentation available on the web pages, as a lot of useful information can be found there.

1. The PROSITE database and ScanProsite

The aim of this exercise is to explore and understand the PROSITE database.

General information in a PROSITE entry

First have a look at the following PROSITE entry: PS50235.

- Is the descriptor PS50235 a PROSITE pattern or a profile? How do you distinguish one from the other?
- Assuming that this descriptor matched your sequence, would you necessarily believe the result? And if it did not match?
- What is the predicted function of proteins matching PS50235? (*Hint* - check the PDOC documentation)
- Is PS50235 related to other PROSITE entries? If yes, are they patterns or profiles? What is the quality of these signatures? (*Hint* - check the PDOC documentation)

Sequence analysis using PROSITE

Analyze the following sequence using ScanProsite:

```
>seq1
MELRVLLCWASLAAALEETLLNTKLETADLKWVTFPQVDGQWEELSGLDEEQHSVRTYEV
CDVQRAPGQAHWLRTGWVPRRGAVHVYATLRFTMLECLSLPRAGRSCKETFTVFYYESDA
DTATALTPAWMENPYIKVDTVAAEHLTRKRPGAEATGKVNVKTLRLGPLSKAGFYLAFQD
QGACMALLSLHLFYKKCAQLTVNLTRFPETVPRELVVPVAGSCVVDAVPAPGPSPSLYCR
EDGQWAEQPVTGCSCAPGFEAAEGNTKCRACAQGTFKPLSGEGSCQPCPANSHSNTIGSA
VCQCRVGYFRARTDPRGAPCTTPPSAPRSVVSRLNGSSLHLEWSAPLESGGREDLTYALR
CRECRPGGSCAPCGGDLTFDPGPRDLVEPWVVVRGLRPDFTYTFEVTALNGVSSLATGPV
PFEPVNVTTDREVPPAVSDIRVTRSSPSSLSLAWAVPRAPSGAVLDYEVKYHEKGAEGPS
SVRFLKTSENRAELRGLKRGASYLVQVRARSEAGYGPFGQEHHSQTQLDESEGWREQLAL
IAGTAVVGVVLVLVVIVVAVLCLRKQSNGREAEYSDKHGQYLIGHGTKVYIDPFTYEDPN
EAVREFAKEIDVSYVKIEEVIGAGEFGEVCRGRLKAPGKKESCVAIKTLKGGYTERQRRE
FLSEASIMGQFEHPNIIRLEGVVTNSMPVMILTEFMENGALDSFLRLNDGQFTVIQLVGM
LRGIASGMRYLAEMSYVHRDLAARNILVNSNLVCKVSDFGLSRFLEENSSDPTYTSSLGG
```

```
KIPIRWTAPEAIAFRKFTSASDAWSYGIVMWEVMSFGERPYWDMSNQDVINAIEQDYRLP
PPPDCPTSLHQLMLDCWQKDRNARPRFPQVVSALDKMIRNPASLKIVARENGGASHPLLD
QRQPHYSAFGSVGEWLRAIKMGRYEESFAAAGFGSFELVSQISAEDLLRIGVTLAGHQKK
ILASVQHMKSQAKPGTPGGTGGPAPQY
```

- What is the domain composition of the protein? What is its function?
- Can you identify potential binding sites using PROSITE? (*Hint* - Move the mouse on the sequences and images in the results page to highlight information)

Now look at this sequence from a patient with a cardio-vascular disease.

```
>seq2
MELRVLLCWASLAAALEETLLNTKLETADLKWVTFPQVDGQWEELSGLDEEQHSVRTYEV
CDVQRAPGQAHWLRTGWVPRRGAVHVYATLRFTMLECLSLPRAGRSCKETFTVFYYESDA
DTATALTPAWMENPYIKVDTVAAEHLTRKRPGAEATGKVNVKTLRLGPLSKAGFYLAFQD
QGACMALLSLHLFYKKCAQLTVNLTRFPETVPRELVVPVAGSCVVDAVPAPGPSPSLYCR
EDGQWAEQPVTGCSCAPGFEAAEGNTKCRACAQGTFKPLSGEGSCQPCPANSHSNTIGSA
VCQCRVGYFRARTDPRGAPCTTPPSAPRSVVSRLNGSSLHLEWSAPLESGGREDLTYALR
CRECRPGGSCAPCGGDLTFDPGPRDLVEPWVVVRGLRPDFTYTFEVTALNGVSSLATGPV
PFEPVNVTTDREVPPAVSDIRVTRSSPSSLSLAWAVPRAPSGAVLDYEVKYHEKGAEGPS
SVRFLKTSENRAELRGLKRGASYLVQVRARSEAGYGPFGQEHHSQTQLDESEGWREQLAL
IAGTAVVGVVLVLVVIVVAVLCLRKQSNGREAEYSDKHGQYLIGHGTKVYIDPFTYEDPN
EAVREFAKEIDVSYVKIEEVIGAGEFGEVCRGRLKAPGKKESCVAISTLKGGYTERQRRE
FLSEASIMGQFEHPNIIRLEGVVTNSMPVMILTEFMENGALDSFLRLNDGQFTVIQLVGM
LRGIASGMRYLAEMSYVHRDLAARNILVNSNLVCKVSDFGLSRFLEENSSDPTYTSSLGG
KIPIRWTAPEAIAFRKFTSASDAWSYGIVMWEVMSFGERPYWDMSNQDVINAIEQDYRLP
PPPDCPTSLHQLMLDCWQKDRNARPRFPQVVSALDKMIRNPASLKIVARENGGASHPLLD
QRQPHYSAFGSVGEWLRAIKMGRYEESFAAAGFGSFELVSQISAEDLLRIGVTLAGHQKK
ILASVQHMKSQAKPGTPGGTGGPAPQY
```

- What is the difference between this sequence and the previous one?
- Can you suggest a possible underlying mechanism for this patients disease?
- Complement your answer using information you can retrieve using UniProt.

Hint - a literature reference.

2. The HAMAP database and HAMAP-Scan

The HAMAP-Scan "Scan" mode can be used to classify protein sequences using HAMAP profiles, while the HAMAP-Scan "Scan & Annotate" mode also provides annotation covering individual sequences and complete proteomes. Both are available here. Complete proteomes can be obtained from UniProtKB in FASTA format as shown using this sample query.

To save time we have annotated a number of proteomes and individual sequences for you. The corresponding results from each of these sequences can be retrieved from the HAMAP-Scan results page using the access codes provided in the following tables.

2a. HAMAP classification and (conditional)

annotation

Individual sequences:

Species name	Taxonomic identifier	HAMAP-Scan access code
Escherichia coli (strain K12)	83333	DLS
Bacillus cereus var. anthracis (strain CI)	637380	OCE

Retrieve the annotations for the individual sequences of *Escherichia coli* (strain K12) and *Bacillus cereus var. anthracis* (strain CI) shown below at the HAMAP-Scan results page. At the same time, copy/paste each of these sequences into the HAMAP-Scan search box here and use the simple "Scan" mode to search all HAMAP profiles in real time for matches to each of the sequences.

```
>Escherichia coli
MLKIFNTLTRQKEEFKPIHAGEVGMYVCGITVYDLCHIGHGRTFVAFDVVARYLRFLGYK
LKYVRNITDIDDKIIKRANENGESFVAMVDRMIAEMHKDFDALNILRPDMEPRATHHIAE
IIELTEQLIAKGHAYVADNGDVMFDVPTDPTYGVLSRQDLDQLQAGARVDVVDDKRNPMD
FVLWKMSKEGEPSWPSPWGAGRPGWHIECSAMNCKQLGNHFDIHGGGSDLMFPHHENEIA
QSTCAHDGQYVNYWMHSGMVMVDREKMSKSLGNFFTVRDVLKYYDAETVRYFLMSGHYRS
QLNYSEENLKQARAALERLYTALRGTDKTVAPAGGEAFEARFIEAMDDDFNTPEAYSVLF
DMAREVNRLKAEDMAAANAMASHLRKLSAVLGLLEQEPEAFLQSGAQADDSEVAEIEALI
QQRLDARKAKDWAAADAARDRLNEMGIVLEDGPQGTTWRRK

>Bacillus cereus var. anthracis
MTIHIYNTLTRQKEEFTPLEENKVKMYVAGPTVYNYIHIGNARPPMVFDTVRRYLEYKGY
DVQYVSNFTDVDDKLIKAANELGEDVPTIADRFVEAYFEDVTALGCKHATVHPRVTENMD
IIIEFIQELVNKGYAYESEGDVYFRTKEFEGYGKLSHQPIADLRHGARIEVGEKKQDPLD
FALWKAAKEGEIFWESPWGQGRPGWHIECSAMARKYLGDTIDIHAGGQDLAFPHHENEIA
QSEALTGKTFARYWMHNGYININNEKMSKSLGNFILVHDIIKQYDPQLIRFFMLSVHYRH
PINFSEELLQSTNNGLERIKTAYGNLKHRMESSTDLTDHNEKWLADLEKFQTAFEEAMND
DFNTANAITELYNVANHANQYLLEEHTSTVVIEAYVKQLETLFDILGLELAQEELLDEEI
EELIQKRIEARKNRDFALSDQIRDDLKDRNIILEDTAQGTRWKRG
```

- Compare the output of the HAMAP-Scan "Scan" and "Scan & Annotate" modes. What information is specific to each?
- Now compare the annotation of the two sequences from the HAMAP-Scan "Scan & Annotate". What differences can you see in the resulting annotation?
- Can you deduce the reasons for these differences? Hint: examine the 'cases' in the HAMAP rule used to produce these annotation (the rule can be identified from the "DR HAMAP" line) as well as the annotation "warnings" provided for each sequence (in the "INTERNAL SECTION"). Can you see why it is necessary to specify a taxonomic node when performing "Scan & Annotate"?

2b. HAMAP proteome annotation - in cooperation with PROSITE

Next, retrieve the annotations for each of the complete proteomes from the HAMAP-Scan results page using the access codes provided below.

Complete proteomes:

Species name	Taxonomic identifier	HAMAP-Scan access code
Escherichia coli (strain K12)	83333	JWH
Buchnera aphidicola subsp. Acyrthosiphon pisum (strain APS)	107806	HYJ
Halopiger xanaduensis (strain DSM 18323 / JCM 14033 / SH-6)	797210	LLN

- What proportion of each proteome has been annotated by HAMAP?
- Can you think of reasons that might explain the variable coverage?

Now examine more closely the the annotation for the complete proteome of Escherichia coli (strain K12). Look in particular at sequence PUR2_ECOLI.

- Which HAMAP rule has annotated this sequence?
- Compare the annotation of this entry to that found in the HAMAP annotation rule. What is the reason for the discrepancies? Hint: examine both the HAMAP rule and the internal section of the annotated entry.
- Given what you deduced about this HAMAP rule in the previous question, why do you think that the HAMAP rule was constructed in this way?

For advanced students

3. Build your own pattern

You are working with a family of proto-oncogene proteins and have identified a potential functional region that is conserved in several related proteins. You have built a multiple sequence alignment of this region from these proteins, and now you would like to identify other proteins having this signature.

- Build a Pattern from the MSA using the PROSITE syntax (you can find it here). *Hint* - you can use weblogo to guide you.

```
Seq1  WFFKGIADKDAERHLLA
Seq2  WFFKNLEQKDAEARLLA
Seq3  WFFKR---KDAERQLLA
Seq4  WFFGTI---DAERQLLA
Seq5  WFFKDIPTKDAERQLLA
Seq6  WYFG----RESERLLLA
Seq7  WYFGKIPLKDAERQLLA
Seq8  WYFGKLRAKDTERLLLL
```

The first thing to do now is the check the quality of your pattern.

- How will you do that? *Hint* - look at the ScanProsite page.

Search the UniProtKB/Swiss-Prot database with your pattern using ScanProsite.

- What can you say about the proteins matching your pattern?

Repeat the exercise with the following sequences:

```
seq1 ERGLAAAR
seq2 DRVSCLIR
seq3 DRLGSGGR
seq4 ERAALILR
seq5 ERIVVTVR
```

- How easy is it to build a good quality pattern? What is the source of any difficulties?

4. MyHits Tutorial

More practicals on how to use MyHits (a SIB resource where you can build your own profiles and HMMs):

- Go to the page http://myhits.vital-it.ch/doc/tutorial-psi.html if you want to do some more practicals on PSI-BLAST.
- Go to this page http://myhits.vital-it.ch/doc/tutorial-domain.html for more practicals in domain hunting.