

Chapter 9 SELF-TEST Protein families and databases

1. In PROSITE, the term PATTERN indicates that the entry describes:

- A. a block
- B. a profile
- C. a regular expression
- D. a fuzzy regular expression

A is wrong – PROSITE doesn't include Blocks

B is wrong – when PROSITE contains profiles, they are referred to with the word MATRIX.

C - correct

D is wrong – PROSITE doesn't include fuzzy regular expressions

2. In PROSITE, the NR lines indicate:

- A. the comment field
- B. the list of true-positive sequences matched by the signature
- C. the statistics of the diagnostic performance of the signature
- D. the non-redundant list of sequences whose 3D structure are known

A is wrong – the comment field is denoted by CC lines.

B is wrong – the list of true positives is indicated on DR lines.

C - correct

D is wrong – structures are listed on 3D lines.

3. When searching the Blocks and PRINTS databases, a match is judged significant if:

- A. a single motif is matched
- B. two motifs are matched
- C. the E-value is above e^{-4}
- D. a combined E-value above a given threshold is reported for a multiple-motif match

A is wrong – a single motif is seldom significant.

B is wrong – two motifs are rarely significant, unless the signature only contains 2 motifs.

C is wrong – E-values vary with the database size, so a fixed value cannot determine absolutely whether a match is significant or not.

D – only when there are several motifs matched, and a significant E-value is reported for the multiple match, is the complete match considered significant.

4. TrEMBL is:

- A. an automatically annotated composite protein sequence database
- B. an automatically annotated supplement to the EMBL database
- C. an automatically annotated supplement to the InterPro database
- D. a translation of coding sequences in the EMBL database

A is wrong – TrEMBL is automatically annotated, but is not a composite database.

B is wrong – TrEMBL is a protein sequence database, not a nucleotide database.

C is wrong – TrEMBL is not a protein family database.

D - correct

5. UniProt is:

- A. the universal protein sequence database derived from SWISS-PROT and TrEMBL
- B. the universal protein resource derived from SWISS-PROT, TrEMBL and PIR-PSD
- C. the universal protein family resource
- D. the universal protein structure database

A is wrong – UniProt includes PIR-PSD.

B - correct

C is wrong – UniProt stores protein sequences, not protein families.

D is wrong – UniProt is not a protein structure database.

6. InterPro is:

- A. an integrated protein family database
- B. an integrated protein sequence database
- C. an integrated protein structure database
- D. an integrated protein interaction database

A - correct

B is wrong – InterPro is a protein family database, not a sequence database.

C is wrong – InterPro is not a protein structure database.

D is wrong – InterPro is not a protein interaction database.

7. In a sequence database of a given size, which of the following expressions is likely to retrieve more matches:

- A. D-A-V-I-D
- B. [DE]-A-V-I-[DE]
- C. [DE]-[AVILM]-X-E
- D. D-A-V-E

A is wrong – the expression is short, and specific, but is longer than D, so will find less matches.

B is wrong – the expression is the same length as A, but is less specific, so will find more matches.

C – because it is shorter and less specific than the others

D is wrong – the expression is shorter than A and B, so will find more matches.

8. Which of the regexs below is not compatible with the following motif:

```
PIFMIPAFYFTWIEMQCS  
PIFMIPAFYFSWIELQGS  
PIFMVPAFYFSWIQMAAS
```

PLMALPAFYFSWWSLVTS
PLMALPAYYFSWWHLKTS
PLVTIGAFFFSWIDL SYS

- A. P-[IL]-[FMV]-X-[IVL]-[PG]-A-[FY]2-F-[TS]-W-[IW]-X-[ML]-X2-S
- B. P-[IVL]-[FMV]-[MAT]-[IVL]-[PG]-A-[FY]2-F-[TS]-W-[IW]-X-[ML]-X2-S
- C. P-[IL]-[FMV]-X-[IVL]-[PG]-A-[FY]4-[TS]-W-[IW]-X-[ML]-X2-S
- D. P-[IVL]-X2-[IVL]-[PG]-A-[FY]2-F-[TS]-W-[IW]-X-[ML]-X2-S

C – all the other options are permutations that are accounted for by the listed sequences; C is a single residue longer than the listed sequences and hence would not be matched by any of them.

9. Which of the following groupings is not compatible with the traditional Venn diagram of overlapping amino acid physicochemical properties:

- A. FYWH, AVILMP, DEQN, KR, STCG
- B. FM, YWH, AVILP, DE, KR, STQN, CG
- C. FYW, AVILM, DE, KRH, STNQ, C, PG
- D. FYW, AVILPST, DEQN, KRH, MCG

The answer is D - all the others group aromatics, or aliphatics, or large hydrophobics, or polars, or basic, or charged, or polar-neutrals, or smalls, or whatever... D lumps Ser and Thr in with the aliphatics, which isn't right by any clustering you can imagine, and throws Met (a large non-polar) in with Cys, which is polar.

10. With knowledge of the physicochemical properties of the amino acids, which of the following hydropathic rankings is unlikely to be correct?:

- A. FILVWAMGTSYQCNPHKEDR
- B. IFVLWMAGCYPTSHENQDKR
- C. IVLRCMAGTSWYPHDNEQFR
- D. FYILMVWCTAPSRHGKQNE

C – because F is hydrophobic but is positioned among the strongly hydrophilic amino acids

11. Which of the sequences below is not compatible with the following regex:

H-[VILM]-G(2)-S-[EDQN]-T-A-[VILM]

- A. HMGGSQ TAM
- B. HVG GSETAV
- C. HIGGSSTAL
- D. HIGGSETAL

C – because the group [EDQN] does not permit Ser.

12. The Midnight Zone is the region of sequence similarity:

- A. above 20% identity
- B. where sequence alignments are not statistically significant, as the same alignment may have arisen by chance
- C. below 20% identity
- D. where sequences fail to be detected by even the most sensitive sequence-based search algorithms

D is correct

13. The Twilight Zone is the region of sequence similarity:

- A. above 50% identity
- B. where sequence alignments are not statistically significant, as the same alignment may have arisen by chance
- C. below 50% identity
- D. where sequences fail to be detected by even the most sensitive sequence-based search algorithms

B – the Twilight Zone is in the region of 10-20% sequence identity, which is the same level of identity that randomly aligned sequences can achieve, and is hence not statistically significant.