

## Math 5233 Regular expression worksheet

Name: \_\_\_\_\_

Regular expressions as used on Prosite for proteins use the following conventions:

- (1) A set of amino acids in square brackets means that any one of those may be present in that position. So [CPW] would mean that a cysteine, proline, or tryptophan could be present in that position.
- (2) A hyphen simply separates positions - they are otherwise meaningless. In regular expressions in programming languages, such redundant features are absent.
- (3) Curly brackets mean that anything but those residues (amino acids) can be present. So {P} would mean anything but a proline there.
- (4) An x means any residue is accepted.
- (5) Numbers in parentheses indicate multiple positions. So x(3) means any three residues. A pair of numbers gives a range, so C(3,4) means that there could be 3 to 4 cysteines there.

```
SAVGTGGMKTKEAAEKA
ARHGSGGMVTKLMAARIA
TGISRGGMITKIRAAQRA
GDFATGGIVTKLIAADFL
NFKGVGGMRTKIKAAKIC
DPYSSGGMISKIEAGKIA
```

- (1) The above sequences are all from a family of glutamate-5 kinases. What are possible regular expressions that would include all these sequences? Which one would give the greatest specificity?
  - (a) [ADGNSTV]-x(2)-[GAS]-[RST]-G(2)-[IM]-x-[ST]-K-[LI]-x-A-[AG]-x(2)-[ALC]
  - (b) [ADGNST]-x(4)-G(2)-[IM]-x-[ST]-K-[LI]-x-A-[AG]-x(2)-[ALC]
  - (c) [ADGNST]-x(2)-[GAS]-x-G(2)-[IM]-x-[ST]-K-[LI]-x-A-[AG]-x(2)-[ALC]
  - (d) [ADGNST]-x(2)-[GAS]-x-G(2)-[IM]-x-[ST]-K-[LI]-x-A-[AG]-[EDKQR]-x-[ALC]

(2) Construct a regular expression pattern for the following block of sequences:

```
GFNIVGYGCTTCIG
GARTEMPGCSLCMG
GFNLVGFGCTTCIG
GGMVLANACGPCIG
GFYLSGFGCTTCIG
GFEWRQSGCSMCLA
GVTLATPGCGPCLG
GALVCNPCCGPCLG
```