COMP 364
Lecture 24
Wednesday, March 5th

Announcements:

- Quiz 1 graded
- HW 2 not yet graded
- HW 3 is supposed to be posted completely by Friday
  - Should be due **March 14th** (late by march 17th)
- Quiz 2 **–** material includes everything from Quiz 1 until the next lecture
  - Arrays, Hashes, Pattern Matching.
  - Will take place on **March 19th** (Notes are allowed)
  - Material from Quiz 1 won't be explicitly tested on.
- Today: Pattern Matching:
  - Anchors, Getting "all" matches, Substitution & Translation

Note:

$str =~ /./ => period matches any character except new line
unless : *$str =~ /./s* => s at the end will allow matching of ANY character including new line

Problem: *Does an amino acid sequence begin with a certain pattern (for example AXV) and end with another pattern(e.g. YYD)?*

`$protSeq =~ /AXV.*YYD/` => Is **incorrect!** The match can occur anywhere in the string, and we are interested in the start and the end of the sequence

Anchors:

^ - matches a pattern at the start of a string
$ - matches a pattern at the end of the string

`$protSeq =~ /^AXV.*YYD$/` => Is the correct solution for the above problem.
There are more anchors in the book, but won't be covered in class.

Getting "all" the matches:

@All = $protSeq =~ /A/g;* => @All gets all matches in *$protSeq
g – stands for global.
Sample Code:

```
$protSeq = "ADSASAQHDSAHUAHOY"
@All = $protSeq =~ /A/g;
print "@All\n"; #will print the letter A 5 times

@All = $protSeq =~ /A|H/g;
print "@All\n"; #will print the letter A 5 times and the letter H 3 times
```

Caveats to “all”:

- each successive match must start after previous ends

'EAAB(1)' =~ /AA|AB/g

/A(A|B)/g

Will find the AA but not the AB

Could be solved by searching for AA and AB separately(on a separate line).

- By default * and + take as much as they can”

  @All = 'BAAAAAA BC' =~ /A+/g;

  - A, AA, AAA, AAAA, AAAAA – will all match the above pattern, but in fact only the longest pattern (AAAAA) will be returned. It will return as many A in a row as it can find. *'BABABABC' =~ /(AB*)./g* => will give ABABABC

- This default behavior can be changed by adding a ? after the * or +.

@All = 'BAAAAABC' =~ /A+?/g -> ('A', 'A', 'A', 'A', 'A') an array that contains a single A five times

@ALL = “BABABABC' =~ /(AB+?)./g ->

? -- will match the desired pattern only once(or as few times as possible – zero or one)

AUG (...)*?UAA – finds first start codon & first subsequent stop codon