BNFO 135 Fall 2010 Programming practice problems – 10/21/2010

1. Write a Python script that prints the percentage of A's, C's, G's, and T's in a DNA sequence
2. Write a Python script that generates a random DNA sequence (containing just A's, C's, G's, and T's) of length k. The script should ask the user for the length k at runtime.
3. Write a Python script that checks for duplicate DNA sequences in a given file. The sequences are in FASTA format. Your script should print the names and sequences of the duplicates.
4. Write a Python script that prints the name and of the longest, shortest, and median length sequence from a file of DNA sequences in FASTA format.
5. Write a Python script that takes as input a query DNA sequence from the user and a filename containing several DNA sequences in FASTA format and prints "True" if the query sequence is found in the file and "False" otherwise.
6. Consider the following output from BLAST in the file called blast.out. Write a Python script to convert the alignment in BLAST format shown below into FASTA format.

```
sp      P08684
CP3A4_HUMAN     Cytochrome P450 3A4 (EC 1.14.13.67) (Quinine 3-monooxygenase) 503 AA
(CYPIIIA4) (Nifedipine oxidase) (Cytochrome P450 3A3)
(CYPIIIA3) (HLp) (Taurochenodeoxycholate
6-alpha-hydroxylase) (EC 1.14.13.97) (NF-25) (P450-PCN1)
[CYP3A4] [Homo sapiens (Human)]

Score =  938 bits (2425), Expect = 0.0
Identities = 467/479 (97%), Positives = 467/479 (97%)

Query: 3   YGTHSHGLFKKLGIPGPTPLPFLGNILSYHKGFCMFDMECHKKYGKVWGFYDGQQPVLAI 62
Sbjct: 25  YGTHSHGLFKKLGIPGPTPLPFLGNILSYHKGFCMFDMECHKKYGKVWGFYDGQQPVLAI 84

Query: 63  TDPDMIKTVLVKECYSVFTNRRPFGPVGFMKSAISIAEDEEWKRLRSLLSPTFTSGKLKE 122
Sbjct: 85  TDPDMIKTVLVKECYSVFTNRRPFGPVGFMKSAISIAEDEEWKRLRSLLSPTFTSGKLKE 144

Query: 123 MVPIIAQYGDVLVRNLRREAETGKPVTLKDVFGAYSMDVITSTSFGVNIDSLNNPQDPFV 182
Sbjct: 145 MVPIIAQYGDVLVRNLRREAETGKPVTLKDVFGAYSMDVITSTSFGVNIDSLNNPQDPFV 204

Query: 183 ENTKKXXXXXXXXXXXXXSITVFPFLIPILEVLNICVFPREVTNFLRKSVKRMKESRLEDT 242
Sbjct: 205 ENTKKLLRFDFLDPFFLSITVFPFLIPILEVLNICVFPREVTNFLRKSVKRMKESRLEDT 264

Query: 243 QKHRVDFLQLMIDSQNSKETESHKALSDLELVAQSIIFIFAGYETTSSVLSFIMYELATH 302
Sbjct: 265 QKHRVDFLQLMIDSQNSKETESHKALSDLELVAQSIIFIFAGYETTSSVLSFIMYELATH 324
```

7. Suppose you are given a DNA sequence in a file called dna.fasta and random fragments of this sequence each of length 5 in a file called fragments.fasta. For example dna.fasta may contain

```
>Genome
ACACAGTGATGATTGAGGGGGGAGAGGACACACAGGGATTGAGATGGA
```

   and fragments.fasta may be

```
>P0
GAGGG
>P1
ACAGT
>P2
TTGAG
>P3
AGGAC
```

   Write a Python script that prints the fragment names in the correct order from left to right as they appear in the sequence in dna.fasta. For the above example the output would be

```
P1, P2, P0, P3
```

8. The Phylip alignment format contains the number of sequences and sequence length in the first line followed by name and sequence. For example

```
2 30
Human       ACCAGGGTAAACGTGGACAATCCGAAAATA
Mouse       AAAAGCGTTAACGTAGACAACCCGAAATTA
```

Consider the interleaved format where there are just 10 characters of a sequence per line:

```
2 30
Human       ACCAGGGTAA
Mouse       AAAAGCGTTA

Human       ACGTGGACAA
Mouse       ACGTAGACAA

Human       TCCGAAAATA
Mouse       CCCGAAATTA
```

Write a Python script that converts interleaved format to non-interleaved. Your script should print the interleaved format to the screen.

9.