Problem Set 2: Regular Expressions

Answer the following questions related to performing the requested tasks and then include your answers in an electronic file, save as a pdf, and merge together as a single pdf. Upload your PDF file to the Blackboard Assignment page for Problem Set 2 by 5 pm on the due date.

Part 0

Download shuf_words.txt from the supplementary materials. This plaintext file contains 354984 different English words, one on each line. Show all of your regular expressions below (write them in the corresponding spaces below).

1. Write a regular expression to find all of the words in this file that contain "cake".
2. Write a regular expression to find the words that have between at least two "i"'s consecutively.
3. Write a regular expression to find words that start in a vowel and contain "cat".

Part 1

You have received the following data involving various crustacean samples (download from Blackboard). Each of these samples has various data associated with it, including species name, collection number, and country in which it was collected. You want your data formatted in the following fashion:

```
>Genus_species_sample number_Country
nucleotide sequence
```

Requirements:

- Genus is represented by the first letter of the genus followed by a period.
- The species is all lower case.
- The sample number contains only digits.
- Country is represented by the first three letters of the Country in all caps.
- The next line consists of sequence data, which can be uppercased or lowercased.

Unfortunately, not everybody who collected data formatted properly. It is up to you, budding bioinformatician, to format the data using regular expressions and your text editors! Please write down the various regular expressions you will use to reformat the data, explaining the various components thereof. Also, attach the modified file to your Blackboard submission so I can review it. It is simpler and possible, but not mandatory, to do all of the requested transformations in a single line.

Part 2

The program you are using requires you to have a list of the collection number and collection site, separated by a colon. For example, the sequence above from Part 1 should become:

```
1659:RUS
```

Now that you've saved your old data, use regular expressions to extract the desired information in the new format and save it as a new file. Then, attach this sample:location list to your homework submission as well.

Reminders

- Be sure that your submission includes three parts: your **written answers to the three sections**, the reformatted file from **Part 1** and the reformatted file from **Part 2**.