# Problem Set 6: Regular Expressions

**The very long lines in this assignment may not display properly in PDF format. See the Markdown GitHub page to see the whole line for long lines of text in this assignment.** 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 assignment files, including your reformatted FASTA files, to the Blackboard Assignment page for Problem Set 6 by 11:59 PM on the due date.

## Part 0

Download shuf_words.txt from the supplementary materials (on Blackboard—sorry for the back and forth between Blackboard and GitHub!). 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

In this part, you will download `Midori_UNIQUE_srRNA_subset.fasta` from Blackboard which is a subset of 562 sequences from a publicly available database of DNA (MIDORI, Machida et al. 2017) that codes for a ribosomal RNA subunit (srRNA). The data are in a FASTA format. The beginning of a typical sequence may look like this:

```
>AB000667.1.2469.3417   root;Eukaryota;Chordata;Actinopteri;Pleuronectiformes;Paralichthyida
CAAAGGCTTGGTCCTGACTTTACTGTCGACTCTAACTAGACTTACACATGCAAGTATCCG
CCCCCCTGTGAGAATGCCCATAACGCCCTGCTCGGGAACAAGGAGCTGGCATCAGGCACA
...
```

The series of letters, numbers and dots immediately after the `>` sign is the name of the sequence (AB000667.1.2469.3417). The naming scheme is as following: `AB000667.1` is the GenBank accession number and version. You can look up the original sequence in GenBank with that. The next part, `.2469.3417`, denotes the the beginning and end of where the sequence was “cut out” of the original, bigger GenBank entry. So this entry in MIDORI is cut out of the original sequences from position 2469 to position 3417.

The next part of the name is separated by a tab (not a space and not multiple spaces). "root;Eukaryota;Chordata;Actinopteri;Pleuronectiformes;Paralichthyidae;Paralichthys;Paralichthys olivaceus" is a summary of the taxonomic classification of the organism that the sequence comes from. *Paralichthys olivaceus* is the species name of the olive flounder, which this sequence comes from. *Paralichthys* is the genus. Paralichthyidae is the family. Pleuronectiformes is the order and *etc.* "root" just denotes the beginning of the taxonomic lineage information.

Using regular expressions, reformat the data to the following format:

```
>AccessionNumber.Version SpeciesName
nucleotide sequence
nucleotide sequence, continued
nucleotide sequence, continued
etc...
```

**Requirements:**

- The sequence name is just the Accession number followed by the version number (e.g. AB000667.1)
- This is followed by a space and then the species name. There should be a space between the genus and species name (e.g. Paralichthys olivaceus).

While looking through the data, you might notice that some sequences are formatted slightly differently for whatever reason! For example, I noticed that some sequences have underscores in the numbers that denote where the sequence is cut out of the larger GenBank sequence. For example, here is one:

```
>AB002132.1._1.403  root;Eukaryota;Arthropoda;Malacostraca;Decapoda;Macrophthalmidae;Macroph
```

It is up to you, budding bioinformatician, to format the data in a way that can accomodate these differences between lines 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, in a manner that is clear to someone who might want to replicate what you did. Save your work into a new FASTA file called "part1.fasta" and attach this 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.

More information on the MIDORI curated databases can be found in the following paper: *Machida,R.J. et al. (2017) Data Descriptor: Metazoan mitochondrial gene sequence reference datasets for taxonomic assignment of environmental samples. Sci. Data, 4, 1–7.*

## Part 2

Suppose that a program you want to use requires you to have the sequence headers to be just the name of the sequence, which should be the the species name with the space replaced with an underscore. For example, the first sequence from Part 1 should become just:

```
>Paralichthys_olivaceus
```

Now that you've saved your old data, use new regular expressions to reformat the data (from either your output from Part 1 or the original input file, just specify which you used) into the new format and save it as a new file called "part2.fasta". Then, attach this "part2.fasta" file to your assignment submission as well.

## Reminders

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