# COMP 364: Homework 4

Rfam alignment (25 points)

*Assigned: April, 2013*

*Due on April 16, 2013 by email before 11:59am*

*Submit all solutions in the form of python script files that the TA can run. Zip your solutions up into a compressed file called* `<your name>_hw4.zip` *and email it to* `jerome.waldispuhl@mcgill.ca` *(indicate COMP364 in the subject).*

## 1 RNA folding (10 pts)

Go to `http://rfam.sanger.ac.uk/family/RF00436` and download of the seed alignment of the Rfam family `RF00436` (N.B. You can use the format of your choice).

- Parse this file using Biopython. (2 pts)
- Enumerate all sequences and run the `RNAfold` program on each on them. Do not forget to remove the gaps! (4 pts)
- Retrieve the secondary structures. Print the sequence and secondary structure. (4 pts)

You will need to call the `RNAfold` program from your python script. To do so, you can use the `subprocess` module.

## 2 Sequence logo (13 pts)

Use the same file as in the previous problem.

- Enumerate all columns $i$ in this alignments and calculate the frequency $f(x)$ of each nucleotide and gaps. (3 pts)
- For each column column, compute the value $H_i = -\sum_{x \in \{A,C,G,U,-\}} f_i(x) \cdot \log f_i(x)$ and print it. (5 pts)
- Using matplotlib, plot a bar chart such that: (i) each column represent a column of the alignment, (ii) each nucleotide is represented by a color (blue for `A`, orange for `C`, red for `G` and green for `U`), and (iii) the height of the section of the column associated with a nucleotide is given by $H_i(x) = f_i(x) \cdot \log f_i(x)$. (5 pts)

Visit `http://matplotlib.org/examples/pylab_examples/bar_stacked.html` to find an example of a bar chart (You will need to remove the error bars though).

**N.B.:** The graph you will print is not a sequence logo. It represents the nucleotide uncertainty per position. To build a sequence logo you will need to compute the confidence in each column as $R_i = 2 - H_i$ and the height of each bar will be calculated as $f_i(x) * R_i$. More details can be found at `http://en.wikipedia.org/wiki/Sequence_logo`.

## 3 End of the term (+2 pts)