**Εισαγωγή στην Βιοπληροφορική**

**Χειμερινό εξάμηνο 2017-2018**

**Χατζηγεωργίου Άρτεμις**

**Παπαδομανωλάκης Ελευθέριος 1634**

**Άσκηση 2: Calculating a weight matrix**

Find the **Position Frequency Matrix** and the **most frequent sequence:**

* If argument after perl’s file name isn’t a valid file, error and exit
* Read every row of file:
  + Check validity and find Position Frequent Matrix:
    - If row contains anything apart from ACTG ,error and exit (/[^ATCG]+/).
    - Assume that every row has the same length as the 1st row. Then if one of the rows has different length, error and exit.
    - Build a hash table with 4 keys (“A”,”C”,”T”,”G”) and an array of ‘length(row)’ assigned to each key
      * In each iteration we scan the row and increment the appropriate position in the hash tables array on the key(base) that we found. In the end of all iterations we have the Position Frequent Matrix
* Print the Position Frequent Matrix
* Build most frequent sequence:
  + Iterate the hash table and, for each key’s entry, keep in a separate array the ‘most seen’ base.
    - If one or more bases have the same value make a string ,with ‘|’ as separator ,containing all these values
* Print the most frequent sequence (mfs)
  + ‘|’ seperates bases with the same value, so we keep all of them in the mfs

Print STDOUT in file:

There are two options for this task:

* Redirect the script’s output to a file
  + perl file.pl input > output.txt
* Write the stdout to a file from ‘inside the script’
  + we implement this in the solution given

**OUTPUT:**

**Position Frequency Matrix 10 (PFM)**

**T: {2, 2, 3, 1, 5, 2, 2, 1, 5, 3, 1, 2, 4, 2, 3, 4, 3, 2}**

**G: {2, 3, 3, 3, 1, 2, 3, 3, 1, 1, 6, 2, 3, 2, 3, 1, 1, 4}**

**C: {2, 3, 2, 5, 2, 4, 3, 3, 2, 4, 3, 3, 2, 3, 3, 3, 2, 3}**

**A: {4, 2, 2, 1, 2, 2, 2, 3, 2, 2, 0, 3, 1, 3, 1, 2, 4, 1}**

**Most frequent sequence**

**A G|C T|G C T C G|C G|C|A T C G C|A T C|A T|G|C T A G**