Homework #2
Computer Science for Bioinformatics (140.636)

Remarks

As usual, to turn in the homework you must create a web page that incorporates descriptions of the problem, the code and the solution. Also your script should be in the directory and be executable by the instructor. The homework path is:

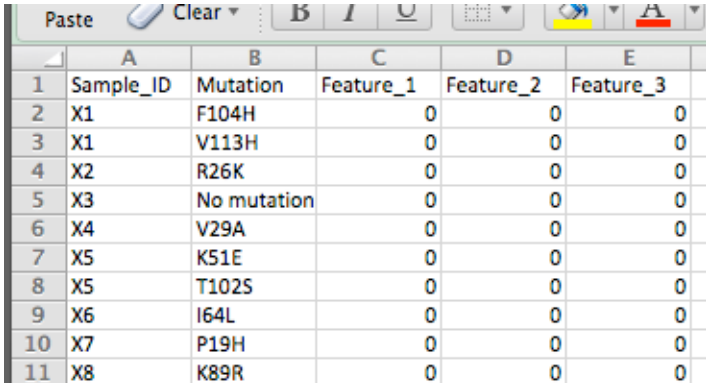
`/users/sph140636/homework/<USERID>/hw2`

To do this exercise you need to understand control structures and regular expressions.

Exercise

One of the most common applications for perl is transforming data from a format provided by a lab and a format that you can process with statistical software, e.g. R. Data provided by a lab often has inconsistent notation missing data, etc. It is rarely in the format needed for process. In this exercise you will take a spreadsheet provided by a lab and convert it into a format that is convenient for analysis in **R**.

The excel spreadsheet can be downloaded from the course web site. The first few rows look like this:



	A	B	C	D	E
1	Sample_ID	Mutation	Feature_1	Feature_2	Feature_3
2	X1	F104H	0	0	0
3	X1	V113H	0	0	0
4	X2	R26K	0	0	0
5	X3	No mutation	0	0	0
6	X4	V29A	0	0	0
7	X5	K51E	0	0	0
8	X5	T102S	0	0	0
9	X6	I64L	0	0	0
10	X7	P19H	0	0	0
11	X8	K89R	0	0	0

Figure 1. Excel spread sheet with data for the exercise

The spreadsheet contains information about mutations found in multiple samples of a protein sequence. Each DNA sample was collected under different conditions. Three binary features (F1, F2, F3) were measured and associated with protein the protein associated with each DNA sequence. The first column in the spreadsheet is a sample identifier. The second column has information about any mutations in the DNA sequence. The last three columns are binary features (0,1) that were measured from the corresponding protein. The notation for the mutation column is **USUALLY** as follows:

'rXs' where 'r' is the single letter code for the encoded amino acid, 'X' is the coordinate of the amino acid and 's' is the type of mutation. There are several type of mutations that are encoded:

```
Substitution:     s = letter from amino acid code
Frame shift       s = ']'
Deletion          s = '#'
Stop              s = '*'
```

Sometimes the lab tossed in other notations. Your task is to write a script that converts the data in the spreadsheet into an "R-friendly" format, i.e. a tab delimited text and a header. Before you start, you will need to export the data in the spreadsheet into a text file.

The columns in the R-friendly text file are nearly the same as in the excel spread sheet. But you want to parse the mutation into three columns. If there are no mutations, the mutation description columns should contain `'na'`. The `residue2` column should contain information about the mutation type. (see the example output below).

sample_ID	residue1	position	residue2	F1	F2	F3
X1	F	104	H	0	0	0
X1	V	113	H	0	0	0
X2	R	26	K	0	0	0
X3	na	na	na	0	0	0
X4	V	29	A	0	0	0
X5	K	51	E	0	0	0
X9	A	72	I	0	0	0
X9	T	102	A	0	0	0
X27	G	131	fshft	0	0	0
R86	I	50	del	1	0	0
C74	L	116	stop	0	1	0