**Project Goal:**

Identifying features of sequence locations in a genome that distinguish two different type of genomic entities. We can work with any genome of our choice.

Here, I have used Armadillo Gene, downloaded from *Ensemble Biomart* website in two Fasta files stored in the Data folder. We use biological understanding to identify and numerically characterize common aspects of different genomic location types. Next, we evaluate the success of our feature characterization by training and testing a two-class machine-learning classifier. We used Sun Grid Engine (SGE) to create a parallel processing pipeline to accomplish this. More details workflow is given in the **Presentation.pdf.**

**How to run this:**

1) Run the entire code as ./RunMe.sh (Detailed execution process written in comments in the **RunMe.sh** script)

2) Each directory has an individual run script.

3) I have deleted directories and file for the first run. Re-running will overwrite some files but it won't affect the execution.

4) It takes about 40-45 seconds to generate results (including splitting files and generating features)

**Result:**

Achieved a classification accuracy of 82-85 %. Using SGE parallel processing reduced the execution time from 2 mins to 42 secs.

Thanks,

~Megha