Yuyue Wang
Bioinfo Proj.1 workflow
11.6.2016

[I: Identify Your Question]

- Species to be studies: Anaerolinea thermophila UNI-1(AT.Uni-1)
- Sequence downloaded from: http://www.ebi.ac.uk/ena/data/view/AP012029&display=txt&expanded=true
- Parsing original data: using Python. Already worked:

```
sequence = 'project_sequence.txt'
test = 'test.txt'
new = 'parsed_sequence.txt'
file = open(sequence,'r')
new_file = open(new,'w')
alldata = file.readlines()
file.close()
for line in alldata:
    if 'SQ ' in line:
        start = alldata.index(line)
        break
raw_sequence_lines = alldata[start+1:-1]
sequence_lines = []
for line in raw_sequence_lines:
    new_line =''
    for i in line:
        if not i.isalpha() and ord(i) != 32:
            break
        new_line += i
    stripped_new_line = new_line.strip()
    sequence_lines.append(stripped_new_line)
new_sequence = '\n'.join(sequence_lines)
new_file.write(new_sequence)
new_file.close()
```

- Genomic features to be studied: dinucleotide and dipeptide frequencies

[II: Further Design]

- How to study the genomic features: compare the
- Hypothesis: there are specific gene segments in AT.Uni-1 that makes it thermophilic compared other species who share close evolutionary relationships with it and is not thermophilic, and thus certain special peptide or dipeptide combinations are produced by these gene segments to make AT.Uni-1 survive high temperature.
- Method to test hypothesis:
    - Steps
        - calculate GC content of, extract dinucleotide and dipeptide pairs from, and calculate their frequencies of AT.Uni-1 sequence.

        - Do the same to species that have close evolutionary relationships with it but is not thermophilic. (needs to download(from EMBL) and parse their sequences befores as what's done on AT.Uni-1 sequence)
        - Create frequency tables for all the species.
        - Compare the difference of dinucleotide and dipeptide frequencies between AT.Uni-1 and its relatives, and see whether there lies a significant distinction.
    - Data type to store data: multi-dimensional array(vector in R?)(will try using R)
    - Statistical test:
        - Pairwise T-test: test how AT.Uni-1 sequence is different from each of its relatives

[III: Report]

- plot(s)
    - dinucleotide and dipeptide pairs versus frequencies in different species
    - X-axis: dinucleotide/dipeptide pair names
    - Y-axis: frequencies
- table(s)
    - Specific gene segment comparison between AT.Uni-1 and different species
    - Columns: name of sequences
    - Rows: gene segments
- Text
    - The meaning of the specific gene segments(whether it points to the thermophilic property of AT.Uni-1)
    - Other possible explanations for the observed different gene
    - Other aspects to be considered when comparing two species
    - Any other detailed explanations and discussions