Guardant Health Bioinformatics Exercise

Question 1

Problem Statement

Given a list of words, print the frequency of each word.

Problem Specification

1. Input is a Python list.
2. Output shall be printed and sorted by *descending order* of frequency of each word, followed by *lexicographically* for each word.
3. Consider possible error modes for your program (i.e program will appropriately exit with error message if input format is incorrect, etc).
4. Discuss examples of possible use cases you shall test.

Sample Input

```
["cat", "dog", "mouse", "cat"]
```

Sample Output

```
cat    2

dog    1

mouse  1
```

Question 2

Problem Statement

Given a collection of at most 10 symbols defining an ordered alphabet, and a positive integer n (n≤10), return all strings of length n that can be formed from the alphabet, ordered lexicographically (use the standard order of symbols in the English alphabet).

Problem Specification

1. Alphabet A has a predetermined order, such as English alphabet organized as (A,B,C,…,Z).
2. Given two strings s and t have same length n, we say that s precedes t in lexicographic order.
3. Discuss possible use cases to test this program.

Sample Dataset

```
A C G T

2
```

Sample Output

```
AA

AC

AG

AT

CA

CC

CG

CT

GA

GC

GG

GT

TA

TC

TG

TT
```

Question 3

Problem Statement

Given an input VCF file with correct format which has unique variants in each line, print a dictionary of key-value pairs where keys are chromosome-positions of each variant and values are "List" of ref_base>alt_base (See example below under the Sample Dataset section).

Problem Specification

1. Variants in VCF are as per VCF v4.2 specification.
2. There can be multiple variants at the same chromosome-position.
3. Considering error handling is a bonus.
4. Discuss basic use cases to test the program.

Sample Dataset – VCF input file with the following content

```
##fileformat=VCFv4.2

##fileDate=20090805

##source=XYZ

##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta ##contig=
<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,s
pecies="Homo sapiens",taxonomy=x> ##phasing=partial

##INFO=<ID=TYPE,Number=1,Type=String,Description="Variant type (snv/indel/
mnv/complex)">

##INFO=<ID=GENE,Number=1,Type=String,Description="Gene symbol">

#CHROM POS    ID      REF    ALT    QUAL    FILTER INFO   FORMAT example1

1      11217286       .      G      A       .      PASS   TYPE=snv;GENE=MT
OR     AF     0.05

1      11217286       .      G      T       .      PASS   TYPE=snv;GENE=MT
OR     AF     0.45

1      11217287       .      T      A       .      PASS   TYPE=snv;GENE=MT
OR     AF     0.15

1      11217288       .      C      G       .      PASS   TYPE=snv;GENE=MT
OR     AF     0.25
```

Sample Output

```
{

       "1-11217286": ["G>A", "G>T"],

       "1-11217287": ["T>A"],

       "1-11217288": ["C>G"]

}
```

Note: Third party libraries can used for parsing VCF files, but not to solve the main problem in question itself. Correctness is the most important feature, followed by good software engineering practices with an eye towards efficiency and generality. Please state what are your key strengths and/or weaknesses of your implementation.