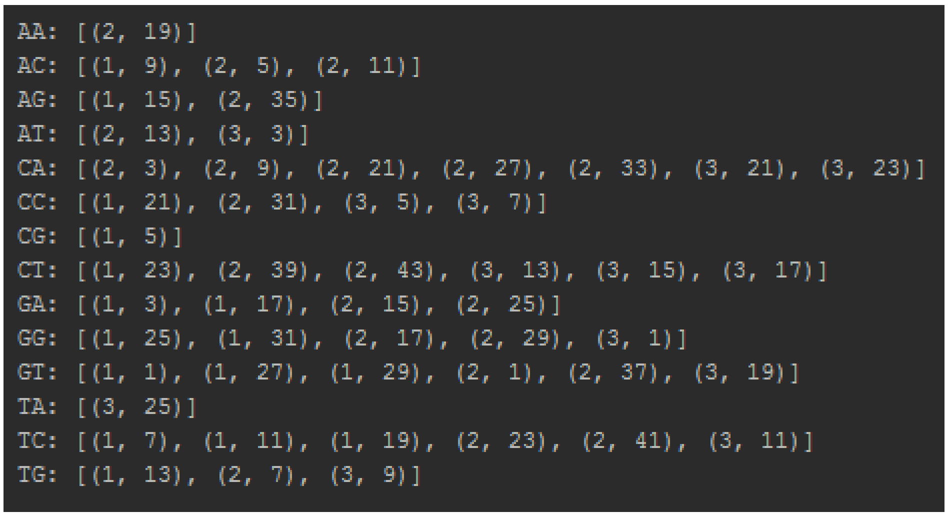
Assignment 2. Retrieval and Mapping.

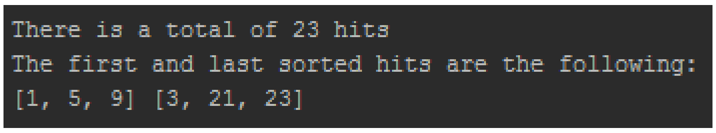
1. Build a hash table for the provided list of sequences "seqs" (same as in the paper). Print the table.



```
AA: [(2, 19)]
AC: [(1, 9), (2, 5), (2, 11)]
AG: [(1, 15), (2, 35)]
AT: [(2, 13), (3, 3)]
CA: [(2, 3), (2, 9), (2, 21), (2, 27), (2, 33), (3, 21), (3, 23)]
CC: [(1, 21), (2, 31), (3, 5), (3, 7)]
CG: [(1, 5)]
CT: [(1, 23), (2, 39), (2, 43), (3, 13), (3, 15), (3, 17)]
GA: [(1, 3), (1, 17), (2, 15), (2, 25)]
GG: [(1, 25), (1, 31), (2, 17), (2, 29), (3, 1)]
GT: [(1, 1), (1, 27), (1, 29), (2, 1), (2, 37), (3, 19)]
TA: [(3, 25)]
TC: [(1, 7), (1, 11), (1, 19), (2, 23), (2, 41), (3, 11)]
TG: [(1, 13), (2, 7), (3, 9)]
```

This is the hash table generated by the 3 sequences provided as database reference using k = 2. Note that those non-existing 2-mers in the sequences aren't stored in the hash-table as keys. (GC and TT don't seem to appear).

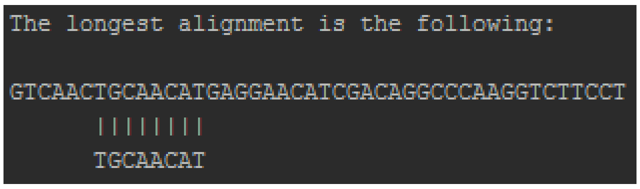
2. Calculate the sorted list of hits for the provided query sequence. Each hit should be a tuple of (index, shift, offset). Print the number of hits and the first and last hit in your list.



```
There is a total of 23 hits
The first and last sorted hits are the following:
[1, 5, 9] [3, 21, 23]
```

This seems to be in concordance with the example provided by the SSAHA algorithm paper.

3. Calculate the maximum match for the provided query sequence, and print the alignment.



```
The longest alignment is the following:

GTCAACTGCAACATGAGGAACATCGACAGGCCCAAGGTCTTCCT
      ||||||||
      TGCAACAT
```

4. DATA: http://www.bioinformatics.nl/courses/BIF-31306/TAIR10.fasta Write a fasta parser to read in the Arabidopsis thaliana genome. How many sequences does the genome file contain and what is the total sequence length? Does this match your expectation on the Arabidopis genome? NOTE: parsing the fasta file should only take a few seconds, otherwise you are doing something very inefficient.

The genome file contains up to 5 sequences, one for each of the chromosomes of *Arabidopsis thaliana.* The total length of the sequences roughly accounts for 119 Mb, which is not that far from the estimated ~135 Mb in the literature.

5. Build a hash table for the Arabidopis chromosomes. Use k = 15. How many keys (different 15-mers) does your hash contain?

There seem to be a total of 7531450 unique *15*-mers in the provided *Arabidopsis thaliana* genome.

6. DATA: http://www.bioinformatics.nl/courses/BIF-31306/athal_query.fasta For the query sequences in the FASTA file athal_query.fasta, find the maximum hit(s) in the genome. Report your findings (if any).

Query sequence 1 finds a hit with the chromosome 3 of *Arabidopsis thaliana* which goes on from position 160650 to position 161056, so a total stretch of 406 bp.

There doesn't seem to be a good enough match for query sequences 2 and 3.

7. What is the algorithm trying to optimize?

The only part of the algorithm that is trying to optimize something is when it tries to find the longest alignment. It checks all the spans of equal indexes and shifts and tries to find the longest one.

8. What did you learn about the time complexity of the search part of the algorithm?

For a given number of *W* and *k*, *W* being the total number of *k*-tuples of length *k* in the hash table (M), the search time is O(n*log(n)) for a query of length *n*. The expected amount of entries in M is $\frac{nW}{4^k}$, so the total search time will logically increase with the amount of tuples in the database and decrease when the length *k* of the tuples increases.

9. What is the main difference in time complexity of the SSAHA algorithm and BLAST?

SSAHA is way faster for large databases because the query is scanned and the whole database is hashed, which takes a long computational time but only once, so search time is independent of the database size. BLAST hashes the query and scans the database, so the search time is directly related to the database size, which gets quite problematic when the database is rather big.

10. How many k-mers with an overlap of p bases does a reference sequence of length m contain?

A reference sequence of length *m* contains $\frac{m-k}{k-p}$ + 1 *k*-mers with an overlap of *p* bases.