Data Structure and Algorithm Design Fall 2014

Project 3

Xiao Wei's Problem II

Zheran Fang, 13212010002@fudan.edu.cn

Dec 1, 2014

The story continues…

Xiao Wei’s Problem II

- Xiao Wei majors in biology
- He’s working on a DNA matching project
- The size of the database is HUGE
- He needs your help, *again*

General Information

- Implement a DNA sequence matching program
- In Java
- Handout: Dec 1, 2014
- Available on FTP `PROJECT` directory

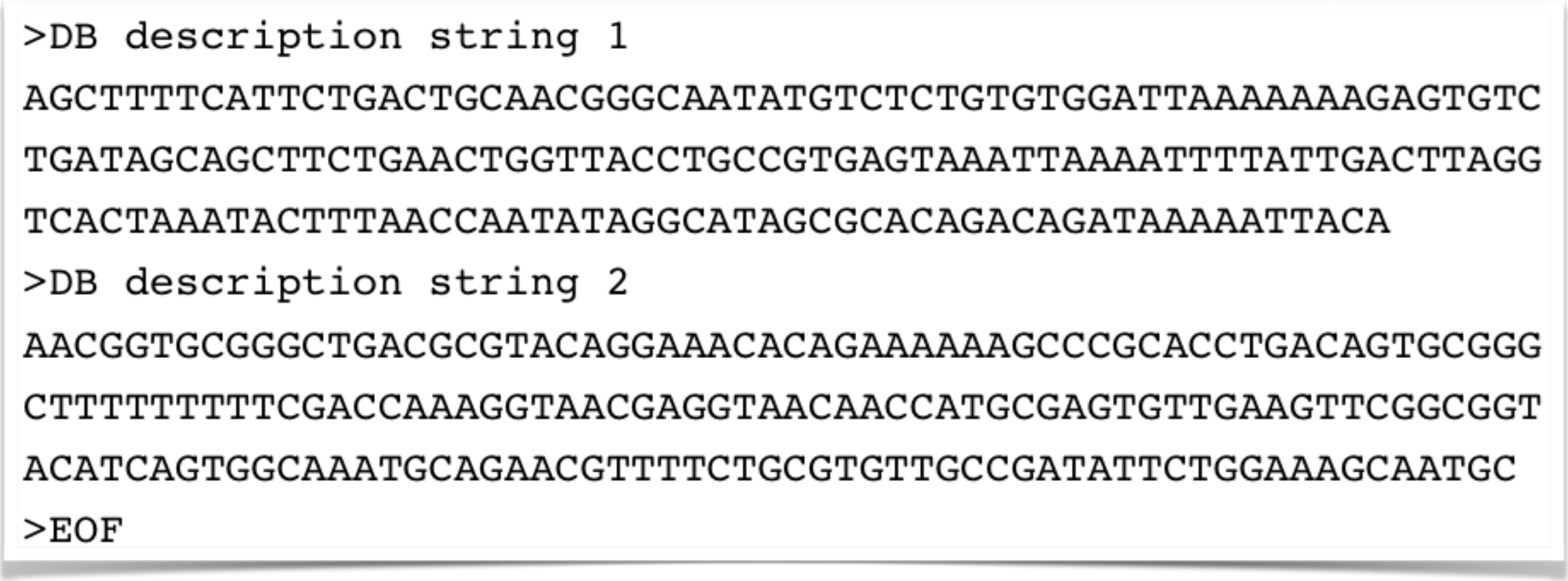
Usage

```
$ java -jar dna_matching.jar db_file query_file output_file
```

Input

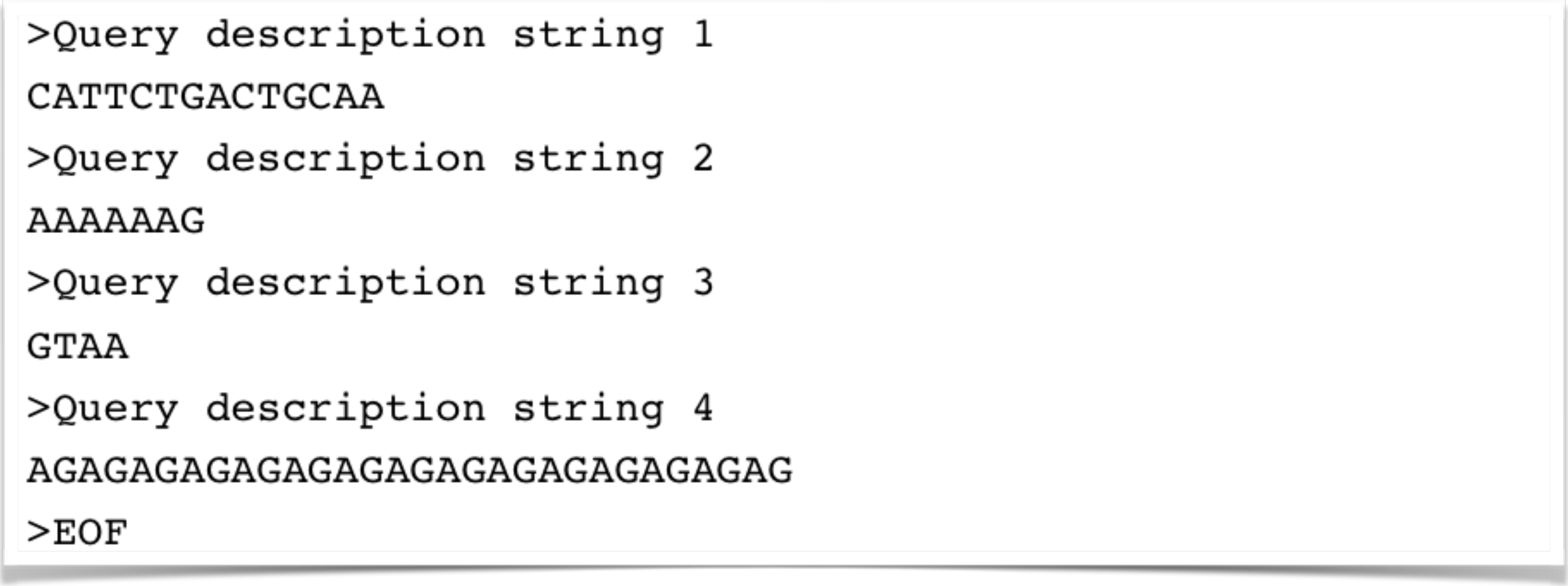
- Sequence format
  - String made up of only four characters: A, C, G and T
  - Prefixed by a line starting with a greater than character (>) followed by a description of the sequence
- Database file
  - Consists of numerous *database sequences*
- Query file
  - Consists of numerous *query sequences*

Sample Database File



```
>DB description string 1
AGCTTTTCATTCTGACTGCAACGGGCAATATGTCTCTGTGTGGATTAAAAAAAGAGTGTC
TGATAGCAGCTTCTGAACTGGTTACCTGCCGTGAGTAAATTAAAATTTTATTGACTTAGG
TCACTAAATACTTTAACCAATATAGGCATAGCGCACAGACAGATAAAAATTACA
>DB description string 2
AACGGTGCGGGCTGACGCGTACAGGAAACACAGAAAAAAGCCCGCACCTGACAGTGCGGG
CTTTTTTTTTCGACCAAAGGTAACGAGGTAACAACCATGCGAGTGTTGAAGTTCGGCGGT
ACATCAGTGGCAAATGCAGAACGTTTTCTGCGTGTTGCCGATATTCTGGAAAGCAATGC
>EOF
```

Sample Query File



```
>Query description string 1
CATTCTGACTGCAA
>Query description string 2
AAAAAAG
>Query description string 3
GTAA
>Query description string 4
AGAGAGAGAGAGAGAGAGAGAGAGAGAGAG
>EOF
```

Output

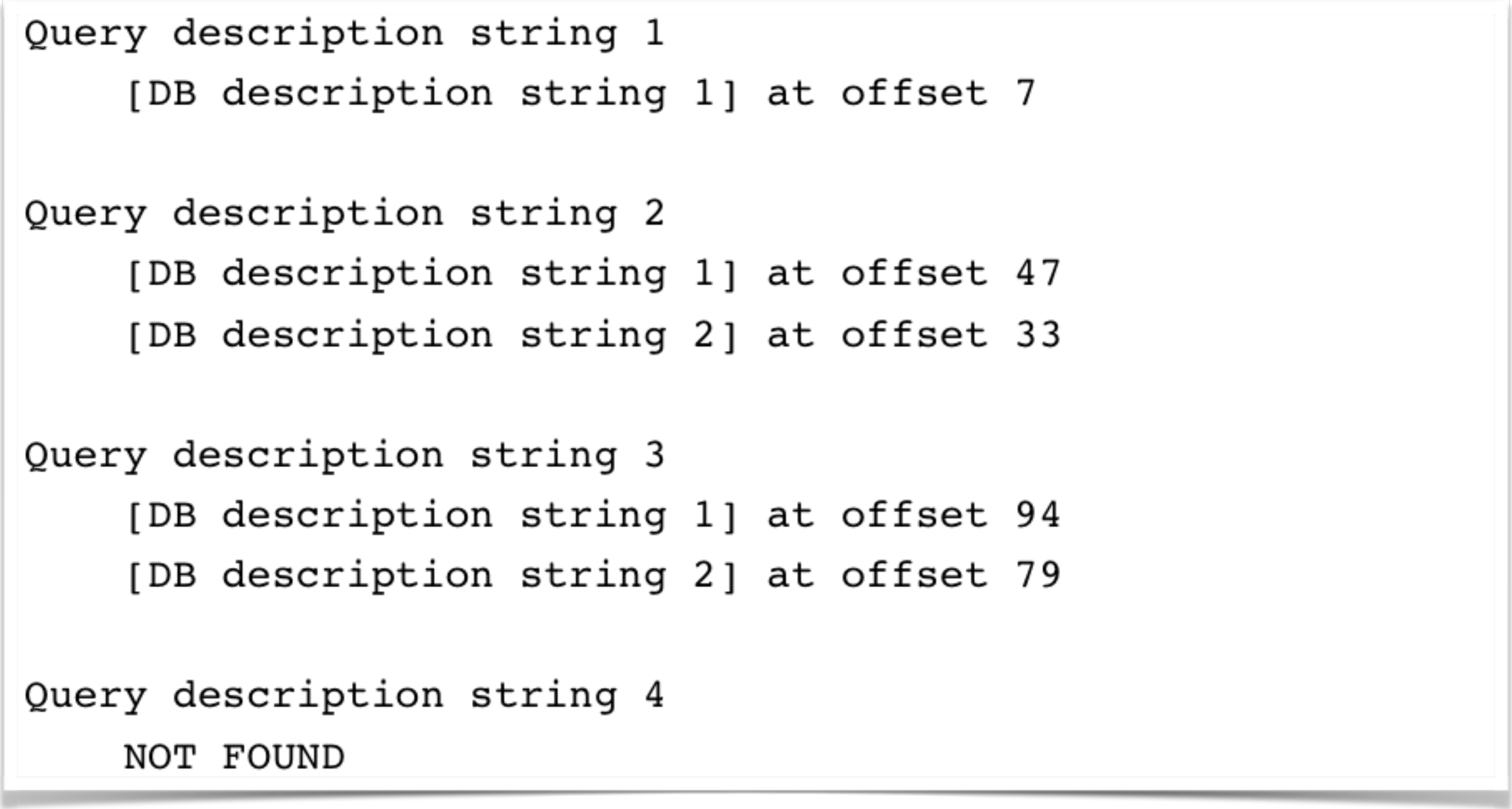
- Report the location of an exact match within any given input DNA sequence for each input search query sequence

- If the query sequence matches within multiple database sequences within the database file, report each result

- If the query sequence matches multiple locations within the same database sequence, report the earliest position of match

- Print `NOT FOUND` otherwise

Sample Output File



```
Query description string 1
    [DB description string 1] at offset 7

Query description string 2
    [DB description string 1] at offset 47
    [DB description string 2] at offset 33

Query description string 3
    [DB description string 1] at offset 94
    [DB description string 2] at offset 79

Query description string 4
    NOT FOUND
```

How to Solve?

- Algorithms
  - Naïve matching
  - Knuth–Morris–Pratt algorithm
  - Boyer–Moore algorithm
  - Rabin–Karp algorithm
- Techniques
  - Multithreading
  - Hashing
  - MapReduce

Grading

- Correctness: 60%
- Time performance: 20%
    - Prerequisite: pass all correctness tests
- Project development document: 15%
- User manual: 5%
- GUI not required

Submission

- Deadline: Dec. 20 2014, 23:59 (GMT+08:00)
- Package your project
  - Source code
  - Executable jar
  - User manual
  - Development document
- Submit to FTP
- Face-to-face interview

Policy

- No cheating
- No late policy

Thanks!

Q&A

Feel free to contact me via E-mail or WeChat