HƯỚNG DẪN THỰC HÀNH

CÁC THUẬT TOÁN ĐỐI SÁNH CHUỖI

(*Keyword: String Searching/String Matching)*

I. Mục tiêu

Sinh viên cài đặt các thuật toán đối sánh chuỗi và ứng dụng nó để giải quyết các bài toán đặt ra.

II. Qui định nộp

- Sinh viên nộp một tập tin nén, có tên là ***<MSSV>.zip*** hoặc ***<MSSV>.rar*** chứa:
  - Thư mục Source: chứa Source code của chương trình.
  - Thư mục Document: chứa tập tin báo cáo như yêu cầu.
  - Thư mục test: chứa các bộ dữ liệu test tự tạo.
- Hạn nộp: xem link trên Moodle.
- Bài giống nhau hay nộp file rác sẽ 0 điểm MÔN HỌC.

III. Nội dung

1. Yêu cầu lập trình:
   Viết chương trình cho phép truy vấn thông tin của chuỗi DNA (Deoxiribo Nucleic Acid) cụ thể trong cơ sở dữ liệu chứa các dãy DNA cho trước bằng cách chọn cài đặt 2 trong số các thuật toán đối sánh chuỗi sau đây: Brute Force, Rabin-Karp, KMP, Boyer Moore.

Cho biết:

- Các dãy DNA được cấu thành từ 4 ký tự A,C,G,T.
- Dữ liệu đầu vào bao gồm 2 tập: Tập cơ sở dữ liệu và Tập các chuỗi truy vấn được mô tả như sau:
  - Tập cơ sở dữ liệu là tập tin văn bản (.txt) chứa N dãy DNA khác nhau ($0<N<2^{32}$). Các dãy DNA có kích thước nhỏ hơn 10^6 byte. Trước mỗi dãy DNA được đánh dấu bởi một dòng mô tả bắt đầu bằng ký tự "?" và sau đó là một chuỗi diễn giải dưới 131 ký tự. Tập tin kết thúc bằng chuỗi ký hiệu "?EOF"

    Ví dụ:

```
?DB string 1
AGCTTTTCATTCTGACTGCAACGGGCAATATGTCTCTGTGTGGATTAAAAAAAGAGTGTCTGATCAGCTT
CTGAACTGGTTACCTGCCGTGAGTAAATTAAAATTTTATTGACTTAGGTCACTAAATACTTTAACCAATA
TAGGCATAGCGCACAGACAGATAAAAATTACA
?DB string 2
```

```
AACGGTGCGGGCTGACGCGTACAGGAAACACAGAAAAAAGCCCGCACCTGACAGTGCGGGCTTTTTTTTT
CGACCAAAGGTAACGAGGTAACAACCATGCGAGTGTTGAAGTTCGGCGGTACATCAGTGGCAAATGCAGA
ACGTTTTCTGCGTGTTGCCGATATTCTGGAAAGCAATGC
?EOF
```

- Tập truy vấn cũng là một tập tin văn bản (.txt) chứa M chuỗi truy vấn khác nhau với những ràng buộc và định dạng tương tự ở tập cơ sở dữ liệu.
  Ví dụ:

```
?Query string 1
CATTCTGACTGCAA
?Query string 2
AAAAAAG
?Query string 3
GTAA
?Query string 4
AGAGAGAGAGAGAGAGAGAGAGAGAGAG
?EOF
```

- Dữ liệu đầu ra của chương trình cho biết kết quả so khớp của từng chuỗi truy vấn với các dãy DNA trong cơ sở dữ liệu. Nếu chuỗi truy vấn trùng khớp nhiều lần trong cùng một dãy DNA thì xuất ra vị trí tìm thấy đầu tiên. Dòng cuối cùng trong tập tin đầu ra cho biết tổng thời gian truy vấn của chương trình. *Lưu ý, vị trí bắt đầu của các chuỗi DNA được tính từ 0.*
  Ví dụ:

```
Query string 1
[DB string 1] at offset 7
Query string 2
[DB string 1] at offset 47
[DB string 2] at offset 33
Query string 3
[DB string 1] at offset 94
[DB string 2] at offset 79
Query string 4
NOT FOUND
Time Total: 10s
```

- Chương trình chạy bằng tham số dòng lệnh được mô tả như sau:
  <Tên chương trình> <Thuật toán> <Tên tập tin cơ sở dữ liệu> <Tên tập tin truy vấn> <Tên tập tin kết quả>
  Trong đó:
  Tên chương trình: là MSSV của tác giả
  Thuật toán: nhận 1 trong 4 giá trị sau:
  - 0: gọi chạy thuật toán Brute Force
  - 1: gọi chạy thuật toán Rabin Karp
  - 2: gọi chạy thuật toán KMP
  - 3: gọi chạy thuật toán Boyer Moore

  Tên tập tin cơ sở dữ liệu, tên tập tin truy vấn và tên tập tin kết quả do người dùng tùy chọn.

2. Yêu cầu về thống kê – báo cáo:
   - Sinh viên tự tạo ra 10 bộ test để test và thống kê hiệu quả các thuật toán mình chọn.

	Test1	*Test2*	*Test3*	*Test4*	*Test5*	*Test6*	*Test7*	*Test8*	*Test9*	*Test10*	*Thời gian TB*
Thuật toán 1											
Thuật toán 2											

Khuyến nghị: sinh viên nên tự viết hàm random ra dãy DNA ngẫu nhiên với các kích thước khác nhau và đủ lớn để test thử chương trình của mình và có thể nộp kèm trong bài nộp của mình.

- Nêu ngắn gọn lý do bạn chọn thuật toán và nhận định của bạn dựa trên kết quả thực tế.
- Nêu rõ phần tự đánh giá của bạn về những công việc bạn đã hoàn thành được so với yêu cầu đề bài.