ĐẠI HỌC THÁI NGUYÊN
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN&TRUYỀN THÔNG

PHẠM THỊ NGA

CÂY QUẢN LÍ ĐOẠN VÀ ỨNG DỤNG

LUẬN VĂN THẠC SĨ KHOA HỌC MÁY TÍNH

Thái Nguyên - 2015

LỜI CAM ĐOAN

Tôi xin cam đoan luận văn này của tự bản thân tôi tìm hiểu, nghiên cứu. Các tài liệu tham khảo được trích dẫn và chú thích đầy đủ. Nếu không đúng tôi xin hoàn toàn chịu trách nhiệm.

Tác giả luận văn

Phạm Thị Nga

LỜI CẢM ƠN

Lời đầu tiên tôi xin được bày tỏ lòng biết ơn chân thành đến Ban Giám Hiệu, các thầy giáo, cô giáo phòng Sau đại học trường Đại học Công Nghệ Thông Tin & Truyền Thông, các thầy giáo ở Viện Công Nghệ Thông Tin đã giảng dạy và tạo mọi điều kiện cho tôi học tập, nghiên cứu và hoàn thành luận văn này.

Đặc biệt, tôi xin bày tỏ sự kính trọng và lòng biết ơn sâu sắc đến PGS. TSKH. Vũ Đình Hòa, người đã tận tình hướng dẫn và giúp đỡ tôi trong suốt quá trình học tập, nghiên cứu và hoàn thành luận văn.

Tôi chân thành cảm ơn các thầy cô tổ Tin học, trường Trung học phổ thông chuyên Lam Sơn, Thanh Hóa, nơi tôi công tác đã tạo điều kiện và hỗ trợ tôi trong suốt thời gian qua.

Tôi cũng xin chân thành cảm ơn người thân, bạn bè đã giúp đỡ và động viên tôi trong suốt thời gian học tập cũng như trong thời gian thực hiện luận văn.

Xin chân thành cảm ơn !

Thanh Hóa, ngày 10 tháng 04 năm 2015

MỤC LỤC

Trang

DANH MỤC CÁC BẢNG

Trang

DANH MỤC CÁC HÌNH

Trang

DANH MỤC CÁC KÍ HIỆU, CHỮ VIẾT TẮT

A	adenine
ADN	Axit deoxyribonucleic
ARN	Axit ribonucleic
BED	Browser Extensible Data
C	cytosine
G	guanine
mARN	messenger ARN
NST	nhiễm sắc thể
PTB	Polypyrimidine Tract-Binding protein
rARN	ribosomal ARN
T	thymine, thymidine
tARN	transfer ARN
U	uracil
UCSC	University of California Santa Cruz

MỞ ĐẦU

1. Lý do chọn đề tài

Từ những năm 2001 trở lại đây, sự tiến bộ về công nghệ và sự phổ cập của các hệ thống phần mềm tiên tiến đã đưa đến những sự thay đổi cách đào tạo chuyên gia trong lĩnh vực tin học. Các kiến thức giải thuật được coi là đỉnh cao trước đây bây giờ đã trở thành "bảng cửu chương" mà ai cũng phải biết và phải thuộc. Những giải thuật ít dùng và phức tạp thì không nhất thiết phải biết hoặc nhớ vì bất cứ ai và lúc nào cũng có thể tra cứu, tìm kiếm chúng trên internet khi cần thiết. Thử thách bây giờ là ở chỗ ta có thể tìm ra những giải pháp hữu hiệu giải quyết một cách có hiệu quả các bài toán, các vấn đề có mô hình toán học đơn giản nhưng có kích thước lớn hay không?

Để đạt được mục đích đó, người lập trình phải tận dụng tối đa khả năng mà phần cứng và hệ điều hành cung cấp, khai thác tối đa khả năng của công cụ lập trình, sử dụng linh hoạt các cấu trúc dữ liệu. Trong đó, cây quản lí đoạn (interval tree) là một cấu trúc dữ liệu quan trọng, có nhiều ứng dụng trong hình học tính toán, truy vấn cơ sở dữ liệu và xử lí tín hiệu.

Bên cạnh đó, tin sinh học là một lĩnh vực mới, giải quyết các bài toán sinh học bằng các phương pháp của khoa học tính toán với nguồn dữ liệu khổng lồ. Việc so sánh các bộ dữ liệu đa dạng di truyền là căn bản để hiểu hệ gen sinh học. Các nhà nghiên cứu phải khám phá nhiều bộ dữ liệu lớn về các đoạn gen (ví dụ như gen, sắp trình tự) để đặt các kết quả thí nghiệm của họ trong một bối cảnh rộng hơn và thực hiện những khám phá mới. Mối quan hệ giữa các tập hợp dữ liệu về gen thường được đo bằng cách xác định các đoạn giao nhau, nghĩa là, chúng chồng lên nhau và do đó chia sẻ một đoạn gen chung. Với những tiến bộ trong công nghệ sắp trình tự ADN, phương pháp

hiệu quả để đo mối quan hệ có ý nghĩa thống kê giữa nhiều bộ tính năng di truyền là rất quan trọng đối với những phát hiện trong tương lai .

Trong khuôn khổ luận văn thạc sĩ, tôi chọn đề tài nghiên cứu: ***"Cây quản lí đoạn và ứng dụng"***, nghiên cứu về cấu trúc dữ liệu cây quản lí đoạn và thực hiện một phương pháp tiếp cận mới, nhanh chóng và linh hoạt để tìm giao giữa các đoạn gen bằng cách sử dụng cấu trúc dữ liệu này.

2. Đối tượng và phạm vi nghiên cứu

Cây quản lí đoạn và ứng dụng để tìm giao các đoạn gen.

3. Những nội dung nghiên cứu chính

Chương 1. Tổng quan về sinh học phân tử, tin sinh học và bài toán tìm giao các đoạn gen

Chương này trình bày một số khái niệm cơ bản của sinh học phân tử, tổng quan về tin sinh học và bài toán tìm giao của các đoạn gen trong sinh học.

Chương 2. Ứng dụng của cây quản lí đoạn để tìm giao các đoạn gen

Chương này trình bày cấu trúc và các thao tác trên cấu trúc dữ liệu cây quản lí đoạn và ứng dụng nó để giải bài toán tìm giao các đoạn gen.

Chương 3. Mã hóa, thử nghiệm chương trình tìm giao các đoạn gen.

4. Phương pháp nghiên cứu

- Phương pháp nghiên cứu lí thuyết: Tổng hợp tài liệu, suy diễn, quy nạp, các phương pháp hình thức,...
- Phương pháp thực nghiệm: xử lí thống kê, đối sánh,...
- Phương pháp trao đổi khoa học, tổng hợp các kết quả của các nhà nghiên cứu liên quan đến lĩnh vực nghiên cứu, lấy ý kiến chuyên gia.