

BỘ GIÁO DỤC VÀ ĐÀO TẠO  
ĐẠI HỌC THÁI NGUYÊN  
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN & TRUYỀN THÔNG

---

**NINH QUANG TRUNG**

**THUẬT TOÁN XÁC ĐỊNH CHA CHUNG GẦN NHẤT CỦA  
HAI NÚT TRONG CÂY ỨNG DỤNG PHÂN TÍCH ĐA DẠNG  
LOÀI VI SINH VẬT**

*Chuyên ngành: Khoa học máy tính*

*Mã số: 60.48.01*

**LUẬN VĂN THẠC SĨ KHOA HỌC MÁY TÍNH**

**Thái Nguyên, năm 2014**

BỘ GIÁO DỤC VÀ ĐÀO TẠO  
ĐẠI HỌC THÁI NGUYÊN  
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN & TRUYỀN THÔNG

---

**NINH QUANG TRUNG**

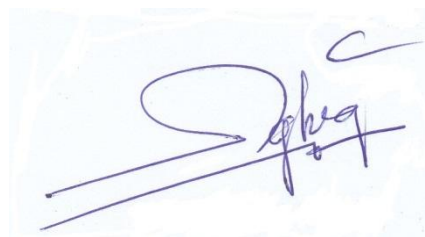
**CÁC PHƯƠNG PHÁP XÁC ĐỊNH CHA CHUNG GẦN  
NHẤT CỦA HAI NÚT TRONG CÂY, ỨNG DỤNG PHÂN  
TÍCH ĐA DẠNG LOÀI VI SINH VẬT**

*Chuyên ngành: Khoa học máy tính*

*Mã số: 60.48.01*

**LUẬN VĂN THẠC SĨ KHOA HỌC MÁY TÍNH**

**Người hướng dẫn khoa học: TS. Nguyễn Cường**



**LỜI CAM ĐOAN**

Tôi xin cam đoan: Luận văn này là công trình nghiên cứu thực sự của cá nhân, được thực hiện dưới sự hướng dẫn khoa học của *Tiến sĩ Nguyễn Cường*. Các số liệu, những kết luận nghiên cứu được trình bày trong luận văn này trung thực và chưa từng được công bố dưới bất cứ hình thức nào.

Tôi xin chịu trách nhiệm về nghiên cứu của mình.

Học viên

**Ninh Quang Trung**

## LỜI CẢM ƠN

Lời đầu tiên, tôi xin chân thành cảm ơn *Tiến sĩ Nguyễn Cường* người đã trực tiếp hướng dẫn tôi hoàn thành luận văn. Với những lời chỉ dẫn, những tài liệu, sự tận tình hướng dẫn và những lời động viên của Thầy đã giúp tôi vượt qua nhiều khó khăn trong quá trình thực hiện luận văn này.

Tôi cũng xin cảm ơn quý Thầy (Cô) giảng dạy chương trình cao học “Khoa học máy tính” đã truyền dạy những kiến thức quý báu, những kiến thức này rất hữu ích và giúp tôi nhiều khi thực hiện nghiên cứu.

Xin cảm ơn các quý Thầy (Cô) công tác tại Trường Đại học Công nghệ thông tin và truyền thông – Đại học Thái Nguyên đã tạo điều kiện cho tôi được tham gia và hoàn thành khóa học.

Tôi xin chân thành cảm ơn.

Học viên

**Ninh Quang Trung**

## MỤC LỤC

LỜI CAM ĐOAN .....	1
LỜI CẢM ƠN .....	iv
MỤC LỤC.....	v
DANH MỤC CÁC HÌNH ẢNH .....	vii
DANH MỤC CÁC BẢNG BIỂU .....	viii
DANH MỤC CÁC TỪ VIẾT TẮT-THUẬT NGỮ .....	ix
MỞ ĐẦU.....	1
CHƯƠNG I: LÝ THUYẾT ĐỒ THỊ VÀ CÂY.....	6
1.1 Các khái niệm cơ bản về đồ thị.....	6
1.1.1 Định nghĩa đồ thị (Graph).....	6
1.1.2 Các khái niệm.....	6
1.1.3 Các thuật toán tìm kiếm trên đồ thị.....	8
1.1.4 Độ phức tạp tính toán của BFS và DFS .....	15
1.2 Các khái niệm cơ bản về cây đồ thị .....	15
1.2.1 Định nghĩa và các tính chất cơ bản:.....	15
1.2.2 Một số khái niệm.....	16
CHƯƠNG II: CÁC PHƯƠNG PHÁP XÁC ĐỊNH CHA CHUNG GẦN NHẤT CỦA HAI NÚT TRONG CÂY.....	17
2.1 Giới thiệu bài toán LCA.....	17
2.2 Mối quan hệ giữa LCA và RMQ .....	19
2.3 Các phương pháp tiếp cận.....	42
2.3.1 Bài toán hà tiện .....	43
2.3.2 Một số phương pháp giải bài toán LCA .....	45
2.4 Lựa chọn phương án cài đặt thuật toán cho bài toán LCA .....	515
CHƯƠNG III: KẾT QUẢ CÀI ĐẶT VÀ ĐÁNH GIÁ .....	54
3.1 Cây phân loài và ứng dụng bài toán phân tích đa dạng loài vi sinh vật .....	548

3.2 Cài đặt phần mềm.....	59
3.3 Đánh giá chất lượng dữ liệu trình tự.....	63
3.4 Lắp ráp trình tự.....	65
3.5 Dự đoán gen .....	66
3.6 Phân tích đa dạng loài vi sinh .....	67
KẾT LUẬN .....	70
TÀI LIỆU THAM KHẢO.....	67

## DANH MỤC CÁC HÌNH ẢNH

Hình 1.1: Ví dụ về mô hình đồ thị .....	6
Hình 1.2: Ví dụ về phân loại đồ thị.....	7
Hình 1.3 Ví dụ về thuật toán tìm kiếm DFS .....	9
Hình 1.4 Xác định đỉnh kề trong thuật toán DFS .....	11
Hình 1.5 Đường đi bắt đầu từ A và kết thúc tại G.....	12
Hình 1.6 Bắt đầu từ A nhưng đi theo trình tự tập các cạnh đã thăm .....	12
Hình 1.7 Duyệt các đỉnh trong cây .....	13
Hình 1.8 Ví dụ thuật toán tìm kiếm theo chiều sâu .....	14
Hình 1.9 Cây đồ thị .....	16
Hình 2.1 Vị trí của các phần tử trong bài toán RQM.....	19
Hình 2.2 Ví dụ về bài toán RQM.....	21
Hình 2.3: Cấu trúc cây phân đoạn.....	23
Hình 2.4 Hình cây của thuật toán LCA.....	29
Hình 2.5 Phân chia đoạn trong bài toán LCA.....	30
Hình 2.6 Chuyển từ bài toán LCA về bài toán RQM .....	35
Hình 2.8 Ví dụ đưa vài toán từ RQM về bài toán LCA.....	37
Hình 2.9 Cây tiến hóa.....	43
Hình 3.1 Quy trình phân tích và xử lý dữ liệu .....	55
Hình 3.2 Chất lượng tính theo vị trí trình tự .....	59
Hình 3.3 Chất lượng theo từng đoạn trình tự.....	60
Hình 3.4 Cây phân loài.....	63
Hình 3.5 Biểu đồ thể hiện sự đa dạng sinh vật trong mẫu dữ liệu.....	63
Hình 3.6 Số lượng các đoạn ORF có liên quan đến các quy trình chuyển hóa .....	65

**DANH MỤC CÁC BẢNG BIỂU**

Bảng 1.1 Bảng lập lịch duyệt các đỉnh trong cây .....	14
Bảng 3.1 Kết quả lắp ráp trình tự.....	61
Bảng 3.2 Tổng quan kết quả dự đoán gen.....	61
Bảng 3.4 Đa dạng loài đã được định tên theo từng cấp độ khác nhau.....	64
Bảng 3.5 Kết quả phân loại theo cơ sở dữ liệu KEGG .....	64



## DANH MỤC CÁC TỪ VIẾT TẮT-THUẬT NGỮ

<b>Cụm từ viết tắt</b>	<b>Cụm từ chi tiết</b>
ASCII	American Standard Code for Information Interchange
BFS	Breadth First Search
Bp	Basepair
DFS	Depth – First – Search
DNA	Deoxyribo Nucleic Acid
E	Edges
G	Graph
GA	Genome Analyzer
HGP	Human Genome Project
LCA	Lowest Common Ancestor
MEGAN	MetaGenomeANalyzer
MGA	MetaGeneAnnotator
NCBI	National Center for Biotechnology Information
NGS	Next Genration Sequencing
PCR	Polymerase chain reaction
RMQ	Range-Minimum Query

## MỞ ĐẦU

Lý thuyết đồ thị là ngành khoa học xuất hiện từ rất lâu nhưng lại có nhiều ứng dụng hiện đại. Những ý tưởng cơ bản của nó được đưa ra từ thế kỷ thứ 18 bởi nhà toán học Thụy Sĩ Leonhard Euler. Ông đã dùng đồ thị để giải quyết bài toán cây cầu Königsberg nổi tiếng. Từ đó lý thuyết đồ thị ngày càng khẳng định được vị trí quan trọng của mình trong việc áp dụng để giải quyết bài toán thực tế nhờ vào việc tìm ra ngày càng nhiều của các định lý, công thức và thuật toán.

Đặc biệt trong những năm trở lại đây, cùng với sự ra đời của máy tính điện tử và sự phát triển nhanh chóng của Tin học, Lý thuyết đồ thị càng được quan tâm đến nhiều hơn. Đặc biệt là các thuật toán trên đồ thị đã có nhiều ứng dụng trong nhiều lĩnh vực khác nhau như: Mạng máy tính, Lý thuyết mã, Tối ưu hoá, Kinh tế học, tìm đường đi ngắn nhất, bài toán luồng cực đại, bài toán vận chuyển, bài toán luồng tổng quát và bài toán xác định cha chung gần nhất của hai nút trong cây cũng là một ứng dụng trong lý thuyết đồ thị. Hiện nay, lý thuyết đồ thị là một trong những kiến thức cơ sở của bộ môn khoa học máy tính.

Trong phạm vi một đề tài không thể nói kỹ và nói hết những vấn đề của lý thuyết đồ thị. Luận văn này trình bày lý thuyết đồ thị dưới góc độ khảo sát những thuật toán cơ bản nhất có thể cài đặt được trên máy tính một số ứng dụng của nó.

Trong khuôn khổ luận văn học viên ứng dụng lý thuyết đồ thị để giải bài toán xác định cha chung gần nhất của hai nút trong cây nhằm mục đích phân tích đa dạng loài vi sinh sử dụng công nghệ đọc trình tự thế hệ mới.

Ngày nay thông tin về trình tự gen rất hữu ích trong những nghiên cứu về sinh học phân tử và trong nhiều lĩnh vực ứng dụng như chuẩn đoán, sinh học pháp y, hệ thống sinh học... Quá trình đọc trình tự hay giải trình tự