

**ĐẠI HỌC THÁI NGUYÊN  
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG**

**HÀ THỊ THANH HỒNG**

**PHƯƠNG PHÁP ĐÁNH CHỈ SỐ  
CHO CSDL GEN ĐỂ TĂNG TỐC ĐỘ TÌM KIẾM**

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

**Thái nguyên, 2015**

**ĐẠI HỌC THÁI NGUYÊN**  
**TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG**

**Hà Thị Thanh Hồng**

**PHƯƠNG PHÁP ĐÁNH CHỈ SỐ**  
**CHO CSDL GEN ĐỂ TĂNG TỐC ĐỘ TÌM KIẾM**

Chuyên ngành: **Khoa học máy tính**

Mã số: **60.48.01.01**

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

NGƯỜI HƯỚNG DẪN KHOA HỌC:

***TS. Hoàng Đỗ Thanh Tùng***

**Thái nguyên, 2015**

## **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 Hoàng Đỗ Thanh Tùng.

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

***Hà Thị Thanh Hồng***

## LỜI CẢM ƠN

Đầu tiên tôi xin gửi lời cảm ơn sâu sắc nhất tới TS.Hoàng Đỗ Thanh Tùng. Thầy đã hướng dẫn khoa học, đã tận tình chỉ bảo, giúp đỡ tôi thực hiện luận văn.

Tôi xin cảm ơn các thầy cô Trường Đại học Công nghệ Thông tin và Truyền thông - Đại học Thái Nguyên đã giảng dạy và truyền kiến thức cho tôi.

Tôi xin chân thành cảm ơn Ban giám hiệu trường Cao đẳng Công nghiệp Thực Phẩm và các đồng nghiệp trong khoa công nghệ thông tin đã tạo mọi điều kiện giúp đỡ tôi hoàn thành nhiệm vụ học tập.

Cuối cùng, tôi xin cảm ơn những người thân và các bạn bè chia sẻ, giúp đỡ tôi hoàn thành luận văn này.

Mặc dù đã hết sức cố gắng hoàn thành luận văn với tất cả sự nỗ lực của bản thân, nhưng luận văn vẫn còn những thiếu sót. Kính mong nhận được những ý kiến đóng góp của quý Thầy, Cô và bạn bè đồng nghiệp.

Tôi xin chân thành cảm ơn!

*Việt Trì, ngày 10 tháng 6 năm 2015*

***Hà Thị Thanh Hồng***

## MỤC LỤC

LỜI CẢM ƠN.....	i
LỜI CAM ĐOAN .....	ii
MỤC LỤC.....	iii
DANH MỤC CÁC KÝ HIỆU, CÁC CHỮ VIẾT TẮT .....	v
DANH MỤC BẢNG BIỂU .....	vi
DANH MỤC HÌNH VẼ.....	vii
<b>MỞ ĐẦU .....</b>	<b>1</b>
<b>CHƯƠNG 1: GIỚI THIỆU TIN SINH HỌC VÀ CƠ SỞ DỮ LIỆU GEN.....</b>	<b>4</b>
1.1 Giới thiệu tin sinh học .....	4
1.1.1 Định nghĩa .....	4
1.1.2 Sự phát triển tin sinh học ở Việt Nam.....	5
1.2 Sinh học phân tử .....	8
1.2.1 Axit nucleic và nucleotide.....	9
1.2.2 Protein và axit amin.....	10
1.2.3 GEN là gì?.....	11
1.2.4 Nhiễm sắc thể và hệ GEN .....	14
1.3 Cơ sở dữ liệu GEN.....	15
1.3.1 Cơ sở dữ liệu NCBI.....	16
1.3.2 Cơ sở dữ liệu EMBL/EBI.....	19
1.3.3 Cơ sở dữ liệu DDBJ .....	19
1.4 Định dạng dữ liệu sinh học.....	20
1.4.1 Định dạng dữ liệu sinh học theo chuẩn FASTA .....	20
1.4.2 Định dạng dữ liệu sinh học theo dạng ALN/ClustalW .....	22
1.4.3 GENBank .....	22

1.5 Kết luận chương 1 .....	23
<b>CHƯƠNG 2: PHƯƠNG PHÁP ĐÁNH CHỈ SỐ GEN ĐỂ TĂNG TỐC ĐỘ TÌM KIẾM.....</b>	<b>25</b>
2.1. Giới thiệu .....	25
2.2 Cấu trúc dữ liệu hệ GEN và sự cần thiết của chỉ số .....	27
2.2.1 Cấu trúc dữ liệu hệ GEN .....	27
2.2.2 Sự cần thiết và lợi thế của đánh chỉ số cho tìm kiếm tương đồng GEN.....	29
2.3. Phương pháp đánh chỉ số cho CSDL GEN .....	30
2.4 Phương pháp đánh chỉ số dựa trên sự biến đổi cấu trúc chỉ số .....	31
2.5 Phương pháp đánh chỉ số dựa vào kích thước (Length based index algorithms) .....	31
2.5.1 Thuật toán đánh chỉ số dựa trên kích thước cố định .....	32
2.5.2 Thuật toán đánh chỉ số dựa trên kích thước biến đổi .....	35
2.6 Thuật toán Blast.....	40
2.6.1 Giới thiệu.....	40
2.6.2. Thuật toán.....	41
2.7. Kết luận chương 2.....	45
<b>CHƯƠNG 3: CÀI ĐẶT THỬ NGHIỆM PHƯƠNG PHÁP ĐÁNH CHỈ SỐ CHO CƠ SỞ DỮ LIỆU GEN ĐỂ TĂNG TỐC ĐỘ TÌM KIẾM.....</b>	<b>46</b>
3.1 Bài toán .....	46
3.2. Xây dựng chương trình thử nghiệm .....	47
3.2.1. Chuẩn bị dữ liệu .....	47
3.2.2. Lựa chọn giải pháp.....	49
Thuật toán.....	49
3.2.3. Thiết kế hệ thống.....	50
3.3. Kết luận chương 3.....	57

<b>KẾT LUẬN VÀ KIẾN NGHỊ .....</b>	<b>59</b>
<b>DANH MỤC TÀI LIỆU THAM KHẢO.....</b>	<b>61</b>

**DANH MỤC CÁC KÝ HIỆU, CÁC CHỮ VIẾT TẮT**

<b>Từ viết tắt</b>	<b>Viết đầy đủ</b>
CSDL	Cơ sở dữ liệu
GEN	Genome
DNA	Axit Deoxyribo Nucleic
ARN	Axit Ribo Nuclêic
NCBI	National Center for BioInformatic Information
dbEST	data base of Expressed Sequence Tags
MGC	Mamalian GEN Collection
EBI	European Biotechnology Information
BLAST	Basic Local Alignment Search Tool
EMBL	European Molecular Biology Laboratory
OMIM	Online Mendelian Inheritance in Man
EPO	European Patent Office
ISDC	International Sequence Database Collaboration
MIAME	Minimum Information About a Microarray Experiment
ASD	Alternative Splicing Database
ATD	Alternate Transcript Diversity
IPD	Immuno Polymorphism Database IPD
CIB – DDBJ	Center for Information Biology and DNA Data Bank of Japan



## **DANH MỤC BẢNG BIỂU**

Bảng 1.1. Nhiệm vụ của một số Bộ, ngành về bảo tồn quỹ GEN quốc gia..	7
Bảng 1.2. Kết quả bảo tồn, lưu giữ nguồn GEN sinh vật .....	8
Bảng 1.3. Tên đầy đủ, tên viết tắt của năm loại nucleotide. ....	9
Bảng 2.1. Minh họa tư tưởng chính của thuật toán BLAST .....	41

## DANH MỤC HÌNH VẼ

Hình 1.1. Cấu trúc xoắn kép của một trình tự DNA .....	10
Hình 1.2. Minh họa cấu trúc của một axit amin.....	11
Hình 1.3. Minh họa một đoạn GEN trong cấu trúc DNA .....	12
Hình 1.4. Quá trình tổng hợp Protein từ đoạn DNA.....	13
Hình 1.5. Định dạng chuẩn FASTA dùng để lưu giữ thông tin trình tự DNA .....	21
Hình 1.6. Định dạng FASTA lưu giữ nhiều trình tự DNA (Protein).....	23
Hình 2.1. Cơ chế ánh xạ trình tự.....	28
Hình 2.2. Sơ đồ thuật toán BLAST.....	44
Hình 3.1. Kết quả tìm kiếm hệ GEN người trên NCBI .....	47
Hình 3.2. Cơ sở dữ liệu của NCBI.....	48
Hình 3.3. Cơ sở dữ liệu mô phỏng .....	49
Hình 3.4. Giao diện chính .....	52
Hình 3.5. Thông báo lỗi từ hệ thống BLAST khi không tìm thấy dữ liệu về trình tự truy vấn.....	53
Hình 3.6. Giao diện nhập dữ liệu .....	54
Hình 3.7. Kết quả chạy thuật toán BLAST .....	55