ĐẠI HỌC THÁI NGUYÊN

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

PHẠM THỊ THU

CÁC THUẬT TOÁN PHÂN CỤM DỮ DIỆU VÀ ỨNG DỤNG TRONG PHÂN LOẠI PROTEIN

LUẬN VĂN THAC SĨ KHOA HỌC MÁY TÍNH

Thái Nguyên – 2015

ĐẠI HỌC THÁI NGUYÊN

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

PHẠM THỊ THU

CÁC THUẬT TOÁN PHÂN CỤM DỮ DIỆU VÀ ỨNG DỤNG TRONG PHÂN LOẠI PROTEIN

Chuyên ngành: Khoa học máy tính

Mã số : 60 48 01 01

Người hướng dẫn khoa học

PGS.TS. Đoàn Văn Ban

Thái Nguyên - 2015

LỜI CẢM ƠN

Để hoàn thành chương trình cao học và viết luận văn này, tôi đã nhận được sự hướng dẫn, giúp đỡ và góp ý nhiệt tình của quý thầy cô trường Đại học Công nghệ thông tin và Truyền thông. Đặc biệt là những thầy cô ở Viện công nghệ thông tin Hà Nội đã tận tình dạy bảo cho tôi suốt thời gian học tập tại trường.

Tôi xin gửi lời cảm ơn sâu sắc đến PGS.TS Đoàn Văn Ban đã dành nhiều thời gian và tâm huyết hướng dẫn tôi hoàn thành luận văn này.

Mặc dù tôi đã có nhiều cố gắng hoàn thiện luận văn bằng tất cả năng lực của mình, tuy nhiên không thể tránh khỏi những thiếu sót, rất mong nhận được sự đóng góp quí báu của quí thầy cô và các bạn.

Tôi xin chân thành cảm ơn!

LỜI CAM ĐOAN

Tôi xin cam đoan tất cả các nội dung của luận văn này hoàn toàn được hình thành và phát triển từ quan điểm của chính cá nhân tôi, dưới sự hướng dẫn chỉ bảo của PGS.TS Đoàn Văn Ban. Các số liệu kết quả có được trong luận văn tốt nghiệp là hoàn toàn trung thực.

Học viên

Phạm Thị Thu

BẢNG KÝ HIỆU CÁC CHỮ VIẾT TẮT

Chữ viết tắt	Nghĩa tiếng anh	Nghĩa tiếng việt
KDD	Kownledge Discovery in Database	Khám phá tri thức trong cơ sở dữ liệu
CSDL	Data base	Cơ sở dữ liệu
KPDL		Khai phá dữ liệu
CURE	Clustering Using Representatives	Phân cụm dữ liệu sử dụng điểm đại diện
CLARA	Clustering Large Application	Thuật toán phân cụm ứng dụng lớn
SoT	Self-organizing Trees	Cây tự tổ chức
DNA	DesoxyriboNucleic Acid	Phân tử nucleic acid mang thông tin di truyền mã hóa cho hoạt động sinh trưởng và phát triển của các dạng sống
RNA	RiboNucleic Acid	Là một trong hai loại axít nucleic, là cơ sở di truyền ở cấp độ phân tử.
rRNA	ribosome RNA	Là ARN mã hóa và mang thông tin từ AND
tRNA	transfer RNA	Là RNA vận chuyển
mRNA	messenger RNA	RNA thông tin
SCOP	Structural Classification of Proteins	Phân loại cấu trúc các protein
CATH	**C**lass **A**rchitecture **T**opology **H**omologous superfamily	Phân loại cấu trúc protein với CATH
DDD	Dali Domain Dictionary	Từ điển miền Dali
PDB	Protein Data Bank	Ngân hàng dữ liệu protein
FSSP	Families of Structurally Similar Proteins	Dòng họ protein với cấu trúc tương tự

DANH MỤC HÌNH VẼ

Số hiệu hình & tên hình vẽ Trang

DANH MỤC BẢNG

MỤC LỤC

MỞ ĐẦU

Trong những năm gần đây, cùng với sự phát triển vượt bậc của công nghệ thông tin, khả năng thu thập và lưu trữ thông tin của các hệ thống thông tin không ngừng được nâng cao. Theo đó, lượng thông tin được lưu trữ trên các thiết bị nhớ không ngừng tăng lên.

Khai phá dữ liệu là quá trình khám phá các tri thức mới có ích ở dạng tiềm năng trong nguồn dữ liệu đã có. Quá trình khám phá tri thức là một chuỗi lặp gồm các bước: làm sạch dữ liệu, tích hợp dữ liệu, chọn lựa dữ liệu, đánh giá mẫu, biểu diễn tri thức. Khai phá dữ liệu liên quan đến nhiều lĩnh vực khác nhau như: công nghệ cơ sở dữ liệu, lý thuyết thống kê, học máy, khoa học thông tin, trực quan hóa,...

Vấn đề ứng dụng các kỹ thuật khai phá dữ liệu, phân cụm dữ liệu trong Tin sinh học, một lĩnh vực còn khá mới, đã ra đời, sử dụng các công nghệ của các ngành toán học ứng dụng, tin học, thống kê, khoa học máy tính, trí tuệ nhân tạo, hóa học, sinh học để giải quyết các vấn đề của sinh học. Việc tìm hiểu và nghiên cứu phân loại protein đã nổi lên như một hướng đi mới với những trải nghiệm hướng vào việc khám phá cấu trúc của các phân tử sinh học.

Nghiên cứu và ứng dụng một cách hiệu quả các phương pháp khai phá dữ liệu là vấn đề hấp dẫn, đã và đang thu hút sự quan tâm chẳng những của các nhà nghiên cứu, ứng dụng mà của cả các tổ chức, doanh nghiệp. Do đó, tôi đã chọn đề tài nghiên cứu **" Các thuật toán phân cụm dữ liệu và ứng dụng trong phân loại Protein**"