

**ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH  
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN  
KHOA HỆ THÔNG THÔNG TIN**



**KHÓA LUẬN TỐT NGHIỆP**

**DỰ ĐOÁN CẤU TRÚC  
BẬC CAO CỦA PROTEIN**

**Giáo viên hướng dẫn: ThS. Nguyễn Thị Kim Phụng**

**Sinh viên thực hiện : Nguyễn Ngọc Tiến**

**MSSV: 08520600**

**Lớp : HTTT03**

**Khóa : 2008 - 2012**

**THÀNH PHỐ HỒ CHÍ MINH – THÁNG 09 NĂM 2012**

**ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH  
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN  
KHOA HỆ THÔNG THÔNG TIN**



**KHÓA LUẬN TỐT NGHIỆP**

**DỰ ĐOÁN CẤU TRÚC  
BẬC CAO CỦA PROTEIN**

**Giáo viên hướng dẫn: ThS. Nguyễn Thị Kim Phụng**

**Sinh viên thực hiện : Nguyễn Ngọc Tiên**

**MSSV: 08520600**

**Lớp : HTTT03**

**Khóa : 2008 - 2012**

**THÀNH PHỐ HỒ CHÍ MINH – THÁNG 09 NĂM 2012**

## LỜI CẢM ƠN

Trước hết, em xin chân thành tỏ lòng biết ơn đến các thầy cô trong khoa Hệ Thông Tin, trường Đại Học Công Nghệ Thông Tin, Đại học Quốc gia Tp. Hồ Chí Minh đã tạo điều kiện thuận lợi cho chúng em học tập và thực hiện khóa luận tốt nghiệp này.

Em xin bày tỏ lòng biết ơn sâu sắc đến ThS. Nguyễn Thị Kim Phụng là những người hướng dẫn khóa luận này. Trong suốt thời gian thực hiện khóa luận, Cô đã tận tình hướng dẫn và động viên và giúp đỡ rất nhiều, Cô đã cho em những lời khuyên những đóng góp quý báu.

Xin cho con gửi những lời cảm ơn chân thành đến gia đình, Ba Mẹ vì đã luôn là nguồn động viên to lớn, giúp đỡ con vượt qua những khó khăn trong suốt quá trình làm việc.

Cuối cùng, tôi xin chân thành cảm ơn sự giúp đỡ, động viên, nhận xét, đóng góp ý kiến của các anh chị, bạn bè trong quá trình thực hiện khóa luận này.

TP HCM, ngày 01 tháng 09 năm 2012

Nguyễn Ngọc Tiên

## NHẬN XÉT CỦA GIÁO VIÊN HƯỚNG DẪN

Tp. Hồ Chí Minh, ngày tháng năm 2012

ThS. Nguyễn Thị Kim Phụng

## NHẬN XÉT CỦA GIÁO VIÊN PHẢN BIỆN

Tp. Hồ Chí Minh, ngày tháng năm 2012

ThS. Lê Ngô Thực Vi

## LỜI MỞ ĐẦU

Trong những năm gần đây, với sự phát triển như vũ bão của khoa học và công nghệ, đã tạo ra cơ sở lý luận, vật chất và sự liên kết hỗ trợ lẫn nhau, tác động thúc đẩy sự phát triển của mọi lĩnh vực của đời sống xã hội, đặc biệt là trên lĩnh vực tin học, công nghệ Internet và công nghệ sinh học.

Tin sinh học chính là sự hội tụ, hợp tác của cả ba lĩnh vực công nghệ hàng đầu như: tin học – công nghệ thông tin – công nghệ sinh học, cùng cộng tác với nhau để khám phá thế giới sống.

Thực tế cho thấy, từ khi Tin sinh học ra đời đã thực sự trở thành công cụ nghiên cứu mới, trợ giúp đắc lực và hiệu quả, đẩy nhanh tốc độ nghiên cứu và ứng dụng công nghệ sinh học, chắp cánh cho công nghệ sinh học nói chung và sinh học nói riêng tiến lên một tầm cao mới. Nhờ thành tựu của Tin sinh học, thời gian nghiên cứu được rút ngắn “trước đây bạn phải mất nữa năm trong phòng thí nghiệm thì bây giờ bạn có thể dễ dàng tiết kiệm thời gian chỉ với một buổi chiều trước chiếc máy tính”.

Tin sinh học có rất nhiều ứng dụng, vì thế cơ sở dữ liệu của công nghệ sinh học không chỉ dừng lại ở tập hợp các kết quả nghiên cứu thực nghiệm đơn thuần của các nhà khoa học trên khắp thế giới, mà nó còn bao gồm khả năng khai quát hóa, mô phỏng hóa thành những “đối tượng số” của thế giới sinh học.

Trong nhiều chương trình ứng dụng của Tin sinh học thì chương trình dự đoán cấu trúc bậc cao của protein là một ứng dụng quan trọng nhất của Tin sinh học và con người còn đang cách lời giải rất xa. Protein là hợp chất hữu cơ có ý nghĩa quan trọng bậc nhất trong cơ thể sống, chúng tham gia mọi hoạt động sống trong cơ thể sinh vật, từ việc tham gia xây dựng tế bào, mô, đến tham gia hoạt động xúc tác và nhiều chức năng khác... Dự đoán cấu trúc bậc cao protein nhằm xây dựng cơ sở dữ liệu protein, phục vụ cho việc tìm hiểu chức năng và ý nghĩa của protein, hiểu được bản chất của sự sống từ đó cải tạo môi trường sống.

Khi nghiên cứu thực nghiệm, các nhà khoa học đã phát hiện ra các protein trong cơ thể sống không hoạt động riêng lẻ, mà chúng có sự tương tác với nhau thông qua một mối liên hệ nào đó. Những protein tương đồng với nhau nếu như giữa chúng có mối quan hệ từ một tổ tiên chung. Mỗi quan hệ tổ tiên càng gần thì sự tương đồng càng lớn. Mà để tìm ra cấu trúc protein bằng phương pháp thực nghiệm đòi hỏi tốn nhiều thời gian và công sức. Do đó, nhiệm vụ chính của Tin sinh học là giúp dự đoán cấu trúc protein chưa biết từ những đặc điểm của những protein đã biết.

Từ những ý tưởng trên và tầm quan trọng của việc dự đoán cấu trúc của protein, em chọn đề tài “Dự đoán cấu trúc bậc cao của protein” nhằm xây dựng một chương trình hệ thống dự đoán cấu trúc bậc cao của những protein mới để hỗ trợ cho các nhà sinh học và giúp giảm thiểu thời gian và chi phí trong quá trình phân tích chức năng của protein.

## MỤC LỤC

|   |             |
|---|-------------|
| <b>Lời Mở đầu .....</b>                             | <b>i</b>    |
| <b>Mục lục.....</b>                                 | <b>iii</b>  |
| <b>Danh sách hình ảnh.....</b>                      | <b>vi</b>   |
| <b>Danh sách các bảng .....</b>                     | <b>viii</b> |
| <b>Danh sách các từ viết tắt .....</b>              | <b>ix</b>   |
| <b>Chương 1: Mở đầu .....</b>                       | <b>1</b>    |
| 1.1 Tổng quan.....                                  | 2           |
| 1.2 Bài toán dự đoán cấu trúc bậc cao protein ..... | 11          |
| 1.3 Mô hình hóa sơ lược bài toán.....               | 12          |
| 1.4 Mục tiêu đề tài .....                           | 13          |
| 1.5 Nội dung nghiên cứu.....                        | 13          |
| <b>Chương 2: Tổng quan về protein.....</b>          | <b>15</b>   |
| 2.1 Khái quát chung về protein.....                 | 16          |
| 2.2 Chức năng sinh học của protein.....             | 18          |
| 2.2.1 Tạo cấu trúc .....                            | 18          |
| 2.2.2 Xúc tác sinh học.....                         | 18          |
| 2.2.3 Vận chuyển.....                               | 18          |
| 2.2.4 Vận động .....                                | 18          |
| 2.2.5 Bảo vệ và chống đỡ.....                       | 19          |
| 2.2.6 Truyền xung thần kinh .....                   | 19          |
| 2.2.7 Dự trữ chất dinh dưỡng .....                  | 19          |
| 2.3 Cấu trúc protein.....                           | 19          |
| 2.3.1 Cấu trúc bậc một .....                        | 21          |
| 2.3.2 Cấu trúc bậc hai .....                        | 22          |

---

|  |           |
|--|-----------|
| 2.3.3 Cấu trúc bậc ba .....  | 24        |
| 2.3.4 Cấu trúc bậc bốn .....   | 26        |
| 2.4 Cây phân loại protein .....                                      | 26        |
| 2.4.1 Cây phân loại SCOP .....                                       | 27        |
| 2.4.2 Cây phân loại CATH.....  | 29        |
| 2.5 Định nghĩa cấu trúc domain trong protein.....                    | 31        |
| 2.6 Tài nguyên thông tin cấu trúc protein .....                      | 32        |
| 2.6.1 Thông tin cấu trúc 3D của protein .....                        | 32        |
| 2.6.2 Đọc thông tin từ PDB.....                                      | 34        |
| 2.6.3 Đọc thông tin từ cây phân loại SCOP .....                      | 36        |
| 2.6.4 Đọc thông tin từ cây phân loại CATH .....                      | 37        |
| 2.7 Tổng kết chương 2 .....  | 39        |
| <b>Chương 3: Xây dựng mô hình dự đoán .....</b>                      | <b>40</b> |
| 3.1 Biểu diễn cấu trúc 3D của protein bằng ma trận khoảng cách ..... | 42        |
| 3.2 Biểu diễn cấu trúc 3D của protein bằng ma trận kè.....           | 43        |
| 3.3 Phố của đồ thị.....  | 44        |
| 3.4 Mô hình K-Nearest Neighbors (K-NN) .....                         | 45        |
| 3.4.1 Khái niệm chung .....  | 45        |
| 3.4.2 Cơ sở của phương pháp K-Nearest Neighbors .....                | 46        |
| 3.4.3 Chọn k .....   | 47        |
| 3.4.4 Nhận xét .....   | 47        |
| 3.5 Mô hình Support Vector Machine (SVM) .....                       | 48        |
| 3.5.1 Khái niệm chung .....  | 48        |
| 3.5.2 Cơ sở của phương pháp SVM .....                                | 49        |
| 3.5.3 Vai trò của hàm kernel .....                                   | 53        |
| 3.5.4 Huấn luyện SVM .....   | 53        |
| 3.5.5 Nhận xét .....   | 54        |

---

|  |           |
|--|-----------|
| 3.6 Tổng kết chương 3 .....                                      | 54        |
| <b>Chương 4: Xây dựng hệ thống chương trình .....</b>            | <b>56</b> |
| 4.1 Hệ thống chương trình.....                                   | 57        |
| 4.1.1 Quy trình thực hiện tổng quát.....                         | 58        |
| 4.1.2 Mô hình dự đoán cho cấu trúc protein mới .....             | 59        |
| 4.1.3 Một số màn hình chức năng tiêu biểu .....                  | 61        |
| 4.2 Kết quả thực hiện .....                                      | 66        |
| 4.2.1 Các tham số đánh giá phân lớp.....                         | 66        |
| 4.2.2 Môi trường cài đặt.....                                    | 66        |
| 4.2.3 Mô tả tập dữ liệu huấn luyện và tập dữ liệu kiểm thử ..... | 67        |
| 4.2.4 Kết quả thực hiện kiểm thử .....                           | 68        |
| 4.2.5 Kết quả dự đoán cấu trúc protein mới.....                  | 70        |
| 4.3 Tổng kết chương 4 .....                                      | 71        |
| <b>Chương 5: Kết luận và kiến nghị.....</b>                      | <b>73</b> |
| 5.1 Kết luận.....  | 74        |
| 5.2 Kiến nghị .....  | 75        |
| <b>Phụ lục .....</b>   | <b>76</b> |
| A. Phân loại amino acid .....                                    | 76        |
| B. Một số kết quả dự đoán xuất ra dạng file .....                | 80        |
| B.1 Kết quả kiểm thử protein .....                               | 80        |
| B.2 Kết quả dự đoán protein mới .....                            | 81        |
| <b>Tài liệu tham khảo.....</b>                                   | <b>83</b> |

## DANH SÁCH CÁC HÌNH ẢNH

|  |    |
|--|----|
| Hình 1.1 – Các enzyme serine protease .....  | 4  |
| Hình 1.2 – Cấu trúc của protein REG3A .....  | 6  |
| Hình 1.3 – Cấu trúc của protein IL – 17 .....  | 6  |
| Hình 1.4 – Một trong những cấu trúc của protein HIV – 1 (pdb 3h47).....                              | 8  |
| Hình 1.5 – Các cấp độ của hai hệ thống phân loại CATH và SCOP.....                                   | 9  |
| Hình 1.6 – Mô hình sơ lược về dự đoán cấu trúc protein .....   | 12 |
| Hình 2.1 – Cấu trúc chuỗi polypeptide .....  | 16 |
| Hình 2.2 – Cấu trúc tổng quát của một amino acid .....   | 17 |
| Hình 2.3 – Sự hình thành chuỗi polypeptide.....  | 17 |
| Hình 2.4 – Cấu trúc bốn cấp bậc của cấu trúc protein .....   | 20 |
| Hình 2.5 – Cấu trúc bậc một của protein .....  | 21 |
| Hình 2.6 – Xoắn alpha .....  | 22 |
| Hình 2.7 – Nếp gấp beta.....   | 23 |
| Hình 2.8 – Cấu trúc bậc ba của protein (pdb 1ogp) .....  | 24 |
| Hình 2.9 – Liên kết disulfide trong protein .....  | 25 |
| Hình 2.10 – Cấu trúc bậc bốn của protein.....  | 26 |
| Hình 2.11 – Cấu trúc phân lớp theo kiến trúc phân loại SCOP .....                                    | 28 |
| Hình 2.12 – Cấu trúc phân lớp theo kiến trúc phân loại CATH.....                                     | 31 |
| Hình 2.13 – Hai domain trong protein 1glqA dựa vào cây phân loại SCOP .....                          | 32 |
| Hình 2.14 – Sự tăng trưởng của cơ sở dữ liệu PDB trong những năm qua (PDB Static, 24/07/ 2012) ..... | 33 |
| Hình 2.15 – Định dạng tọa độ 3D của protein 1glq trong PDB .....                                     | 35 |
| Hình 2.16 – Thông tin cây phân loại SCOP cho hai domain của protein 1glqA ..                         | 36 |
| Hình 2.17 – Thông tin cây phân loại CATH cho hai domain của protein 1glqA .                          | 38 |
| Hình 3.1 – Biểu diễn ma trận khoảng cách 2D từ cấu trúc 3D của protein .....                         | 42 |

|  |    |
|--|----|
| Hình 3.2 – Đồ thị vô hướng G .....   | 43 |
| Hình 3.3 – K-Nearest Neighbor với các giá trị của k là quá nhỏ, vừa và quá lớn .         | 47 |
| Hình 3.4 – Hai cách chia không gian vector thành hai nữa riêng biệt.....                 | 48 |
| Hình 3.5 – Mặt siêu phẳng tách các mẫu dương ra khỏi các mẫu âm .....                    | 49 |
| Hình 3.6 – Siêu phẳng tách với khoảng cách lề cực đại .....                              | 50 |
| Hình 3.7 – Chuyển không gian ban đầu vào không gian đặc trưng .....                      | 52 |
| Hình 4.1 – Quy trình thực hiện tổng quát .....   | 58 |
| Hình 4.2 – Mô hình dự đoán cho cấu trúc protein mới.....                                 | 60 |
| Hình 4.3 – Màn hình khởi động chương trình.....  | 61 |
| Hình 4.4 – Màn hình dự đoán cấu trúc protein mới .....                                   | 62 |
| Hình 4.5 – Màn hình kiểm thử .....   | 63 |
| Hình 4.6 – Màn hình máy học SVM.....   | 64 |
| Hình 4.7 – Màn hình học dữ liệu từ PDB, SCOP và CATH.....                                | 65 |
| Hình 4.8 – Kết quả kiểm thử cấu trúc protein theo cây phân loại CATH .....               | 68 |
| Hình 4.9 – Biểu đồ so sánh kết quả kiểm thử trên SCOP và CATH đối với mô hình K-NN ..... | 69 |
| Hình 4.10 – Biểu đồ so sánh kết quả kiểm thử trên SCOP và CATH đối với mô hình SVM ..... | 69 |
| Hình 4.11 – Kết quả dự đoán cấu trúc protein mới theo cây phân loại CATH .....           | 70 |
| Hình 4.12 – Biểu đồ so sánh kết quả giữa hai mô hình SVM và KNN .....                    | 71 |
| Hình A.1 – Công thức cấu tạo các amino acid nhóm I .....                                 | 76 |
| Hình A.2 – Công thức cấu tạo các amino acid nhóm II.....                                 | 77 |
| Hình A.3 – Công thức cấu tạo các amino acid nhóm III .....                               | 77 |
| Hình A.4 – Công thức cấu tạo các amino acid nhóm IV .....                                | 78 |
| Hình A.5 – Công thức cấu tạo các amino acid nhóm V .....                                 | 78 |
| Hình B.1 – Kết quả kiểm thử xuất ra dưới dạng file .....                                 | 80 |
| Hình B.2 – Kết quả dự đoán protein mới xuất ra dưới dạng file.....                       | 82 |

## DANH SÁCH CÁC BẢNG

|   |    |
|---|----|
| Bảng 2.1 – Bảng thống kê hệ thống phân loại SCOP version 1.75 .....   | 29 |
| Bảng 2.2 – Bảng thống kê hệ thống phân loại CATH version 3.4 .....    | 31 |
| Bảng 2.3 – Mô tả chi tiết các cột atom của file protein PDB.....      | 34 |
| Bảng 2.4 – Định dạng thông tin từ cây phân loại CATH .....            | 37 |
| Bảng 4.1 – Bảng thống kê dữ liệu huấn luyện và dữ liệu kiểm thử ..... | 67 |
| Bảng A.1 – Hai mươi amino acid trong protein .....                    | 79 |

## DANH SÁCH CÁC TỪ VIẾT TẮT

|       |   |
|-------|---|
| 3D    | Three – dimensional   |
| AIDS  | Acquired Immunodeficiency Syndrome  |
| CATH  | CATH Protein Structure Classification – Class, Architecture, Topology, Homologous Superfamily |
| DNA   | Deoxyribonucleic acid   |
| KNN   | K – Nearest Neighbors   |
| KKT   | Karush Kuhn Tucker  |
| NMR   | Nuclear Magnetic Resonance  |
| PDB   | Protein Data Bank   |
| SCOP  | Structural Classification of Protein  |
| SMO   | Sequential Minimal Optimization   |
| SSAP  | Sequential Structure Alignment Program  |
| SVM   | Support Vector Machine  |
| wwPDB | Worldwide Protein Data Bank   |

## CHƯƠNG 1

# MỞ ĐẦU

## 1.1 Tổng Quan

Tin sinh học (Bionformations) [5], [6], [7] là lĩnh vực khoa học sử dụng các công nghệ của các ngành toán học, tin học, thống kê, khoa học máy tính, trí tuệ nhân tạo, hóa học và hóa sinh để giải quyết các vấn đề sinh học. Mục tiêu chính của Tin sinh học là phát triển các thuật giải tính toán, thống kê phục vụ phân tích các dữ liệu sinh học thực nghiệm từ đó khám phá ra các kiến thức mới về sinh học.

Các lĩnh vực nghiên cứu chính của Tin sinh học bao gồm bắt cặp trình tự (sequence alignment), bắt cặp cấu trúc protein (protein structural alignment), dự đoán cấu trúc protein (protein structure prediction), dự đoán biểu hiện gene (gene expression), tương tác protein – protein (protein – protein interaction), và mô hình hóa quá trình tiến hóa. Một trong những mối quan tâm chính trong các dự án Tin sinh học chính là trích rút các thông tin hữu ích từ các thông tin hỗn độn được thu thập từ các kỹ thuật sinh học với lưu lượng mức độ lớn.

Dữ liệu sinh học ở đây bao gồm trình tự DNA (deoxyribonucleic acid), protein, tế bào, tương tác protein – protein, các thông tin y học... Tin sinh học đã thực hiện lưu trữ các thông tin về cấu trúc của protein trong cơ sở dữ liệu PDB (Protein Data Bank) [30] bao gồm tất cả các cấu trúc 3D (three – dimensional) đã biết của các đại phân tử sinh học. PDB được thành lập vào năm 1971 bắt đầu với 7 cấu trúc protein đến nay đã lên đến con số 83,266 (truy cập ngày 24/07/2012) dữ liệu protein được cập nhật thường xuyên vào thứ tư hàng tuần. Kho lưu trữ PDB thành lập nhằm mục đích phục vụ cộng đồng trên thế giới, các nhà nghiên cứu, các nhà giáo dục, các sinh viên nghiên cứu sinh học. Việc phân tích dữ liệu trong PDB có thể giúp giải thích các bệnh tật, phát triển các loại thuốc mới hoặc hiểu sự tương tác giữa các protein khác nhau [17].

Do lượng dữ liệu thông tin là rất lớn nên việc xử lý dữ liệu bằng các phương pháp truyền thống không thể khai thác được một cách hiệu quả lượng dữ liệu lớn như vậy. Theo đánh giá của IBM, các phương pháp khai thác thông tin bằng

phương pháp truyền thống chỉ thu được khoảng 80% lượng thông tin từ cơ sở dữ liệu, phần còn lại bao gồm các thông tin khái quát, thông tin có tính quy luật vẫn còn đang tiềm ẩn trong cơ sở dữ liệu. Lượng thông tin này tuy nhỏ nhưng là những thông tin cốt lõi, những tri thức quý giá cần thiết cho tiến trình quyết định.

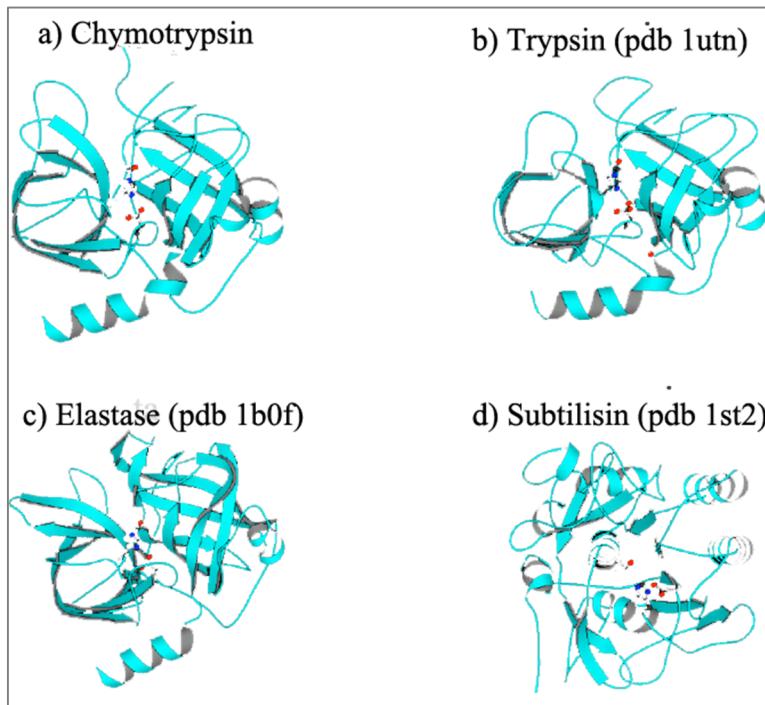
Theo giáo sư Amgad Madkour [7], một trong những ý tưởng quan trọng trong nghiên cứu Tin sinh học là quan điểm tương đồng (homologues). Trong một nhánh hệ gene học (genomic) của Tin sinh học, tính tương đồng được sử dụng để dự đoán cấu trúc của gene: nếu biết trình tự và chức năng của gene A và trình tự này tương đồng với trình tự của gene B chưa biết chức năng thì có thể kết luận rằng A và B có cùng chức năng. Khi nghiên cứu về protein, các nhà khoa học đã chỉ ra rằng: những protein tương đồng với nhau nếu như giữa chúng có mối quan hệ từ một tổ tiên chung. Mỗi quan hệ tổ tiên càng gần thì sự tương đồng càng lớn. Trong nhánh cấu trúc của Tin sinh học, tính tương đồng được dùng để xác định những hợp phần quan trọng trong cấu trúc của protein cũng như tương tác của nó với các protein khác. Với kỹ thuật mô phỏng tính tương đồng (homology modelling), thông tin này được dùng để dự đoán cấu trúc của một protein khi đã biết cấu trúc của một protein khác tương đồng với nó. Hiện tại đây là cách dự đoán cấu trúc protein đáng tin cậy nhất.

Một ví dụ về tính tương đồng giữa HemoGlobin ở người và Leghemo Globin ở các cây họ đậu. Khi nghiên cứu ra Leghemo Globin ở các cây họ đậu. Các nhà khoa học cho rằng nó có cấu trúc và chức năng giống HemoGlobin ở người. Để dự đoán cấu trúc các nhà khoa học đã làm qua các bước theo mô hình tương đồng và nhận thấy rằng Leghemo Globin ở các cây họ đậu có cấu trúc tương đồng với HemoGlobin ở người nên cũng sẽ có chức năng tương đồng là vận chuyển Oxy.

Protein là một đại phân tử hữu cơ có rất nhiều chức năng và đóng vai trò rất lớn trong cơ thể sống [10]. Do đó, dự đoán cấu trúc protein là một trong những nhiệm vụ quan trọng nhằm xây dựng cơ sở dữ liệu protein, phục vụ cho việc tìm hiểu chức

năng và ý nghĩa của protein, hiểu được bản chất của sự sống từ đó cải thiện môi trường sống.

Chức năng của protein chỉ có thể hiểu được trong mối quan hệ với cấu trúc của protein [2], [9], [10]. Điều đó có nghĩa là cấu trúc của protein xác định chức năng sinh hóa của nó. Do mối quan hệ chặt chẽ giữa cấu trúc và chức năng của protein. Dự đoán cấu trúc bậc cao của protein đã trở thành một trong những nhiệm vụ quan trọng nhất trong những năm gần đây. Hình 1.1 là một số protein có cấu trúc tương đồng thì chức năng của chúng cũng tương đồng với nhau.



Hình 1.1 – Các enzyme serine protease

Alan Fersht [15] đưa ra một ví dụ về các enzyme serine protease (hình 1.1). Enzyme Serine Protease trong tất cả các loài có vú, cùng chia sẻ một đặc điểm chung về cấu trúc, thứ tự trong đoạn polypeptide cũng như chức năng. Về cấu trúc, chúng đều có một residue serine rất đặc biệt cho phép phản ứng một chiều với các cấu trúc hóa học có chứa gốc organophosphates, và nhờ đó, enzyme này có khả năng phân cắt các protein tại những residue serine. Những enzyme chính tại tuyến

tuy như: trypsin, chymotrypsin, hay elastase có hoạt động rất giống với nhau trong các hoạt động xúc tác thủy phân các đoạn peptides có tính acid. Từ ví dụ trên của Fersht, ta có thể dễ dàng nhận ra rằng, các cấu trúc serine protease ở loài có vú chắc chắn phải tiến hóa từ một tổ tiên chung trong quá khứ. Các protein có cùng tổ tiên như vậy được gọi là tương đồng.

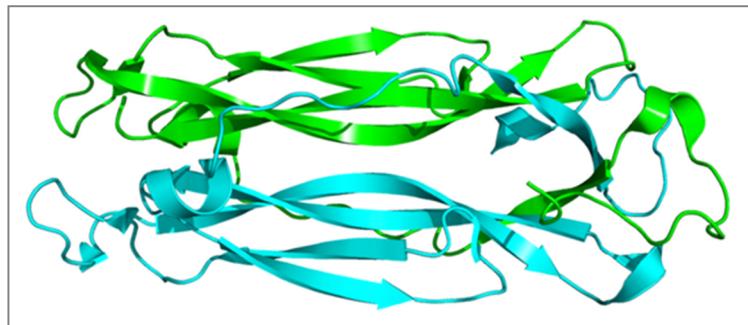
Cấu trúc protein thể hiện qua bốn cấp độ, trong đó cấu trúc bậc ba quy định rõ các hoạt tính chức năng của protein đó và có ý nghĩa rất quan trọng trong việc nghiên cứu các vấn đề về sinh học [7]. Việc tìm hiểu cấu trúc bậc ba giúp rất nhiều cho việc nghiên cứu các chức năng của protein, các vùng hoạt động, các vùng liên kết... và đặc biệt cho việc điều chế và khám phá thuốc, ứng dụng biểu hiện của một protein nào đó. Ví dụ, cấu trúc bậc ba có thể có ích để giải thích các bệnh tật hoặc phát triển các loại thuốc mới. Cấu trúc bậc ba cũng có thể được khai thác để tìm kiếm PDB với các tương tác giữa các protein.

Một ví dụ về ứng dụng trong việc điều chế và khám phá thuốc như là protein REG3A (hình 1.2). Bệnh vảy nến và sẹo là hai bệnh về da vẫn chưa có cách chữa trị hiệu quả và dứt điểm. Theo một báo cáo mới đây [19] được công bố vào ngày 21/06/2012, các nhà khoa học thuộc trường đại học San Diego (Mỹ) đã công bố việc nghiên cứu protein REG3A có thể giúp tìm ra phương pháp chữa trị có hiệu quả các biến chứng này. Bệnh vảy nến là một chứng rối loạn miễn nhiễm của da, khiến các tế bào da phát triển không kiểm soát được, hậu quả sinh ra các vảy sần sùi trên da. Trong khi sẹo xuất hiện sau các tổn thương trên da (như trầy xước, đứt, phỏng...), cả hai chứng này đều gây mất thẩm mỹ khiến người bệnh mất tự tin. Theo giáo sư, tiến sĩ, bác sĩ Richard L. Gallo, trưởng viện bệnh ngoài da của đại học San Diego – người dẫn đầu nhóm nghiên cứu trên cho biết: khi nghiên cứu trên chuột, họ đã phát hiện ra rằng một loại phân tử có tên gọi protein REG3A phát triển rất mạnh ở những vùng da đang phục hồi khi bị tổn thương, vốn không xuất hiện ở những nơi da khỏe mạnh. Các nhà khoa học cũng nhấn mạnh rằng một loại protein

khác có tên là IL-17 (hình 1.3) cũng rất quan trọng, bởi một khi da bị tổn thương, IL-17 sẽ được sinh ra để kích thích protein REG3A xuất hiện. Sự phối hợp nhuần nhuyễn giữa IL-17 và REG3A sẽ giúp cơ thể biết những vùng da nào bị tổn thương để tập trung chữa trị. Với pháp hiện trên, họ tin rằng việc tập trung nghiên cứu REG3A có thể giúp tìm ra biện pháp chữa trị các thương tổn trên da. Bào chế được một phương thuốc giúp kích thích sự sinh ra của protein REG3A sẽ giúp ích việc phát triển của tế bào và cải thiện quá trình lành vết thương.

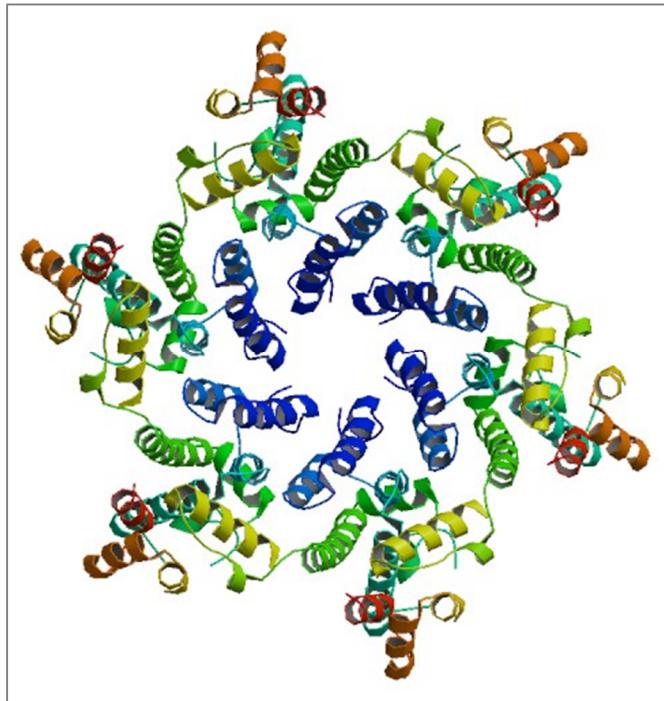


Hình 1.2 – Cấu trúc của protein REG3A  
(Tên đầy đủ là Regenerating Islet-derived Protein 3 alpha, PDB 1uv0)



Hình 1.3 – Cấu trúc của protein IL – 17  
(Tên đầy đủ là Interlaukin – 17, PDB 1jpy)

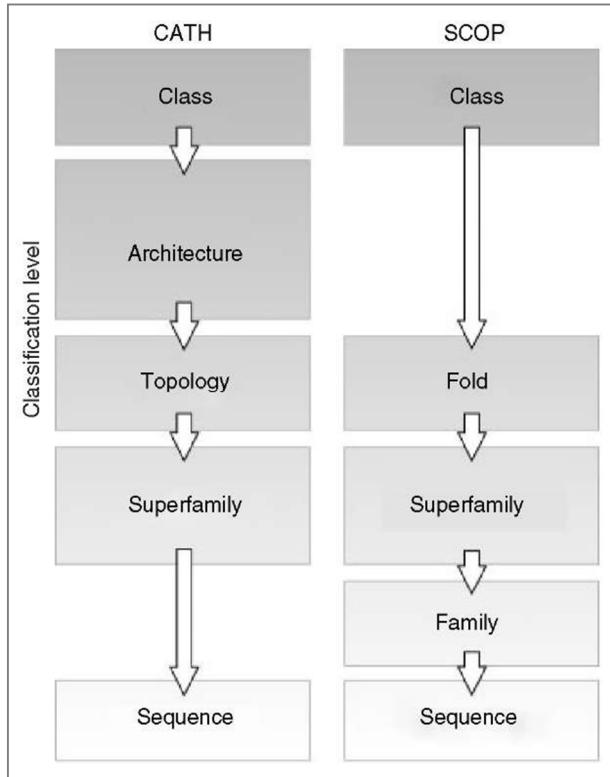
Một ứng dụng khác cũng được xem là rất quan trọng trong việc nghiên cứu chức năng của protein là giải thích các bệnh tật trên cơ thể người. Khóa luận đưa ra một ứng dụng trong thực tế đó là bệnh HIV (Human Immunodeficiency Virus – có nghĩa là virus suy giảm miễn dịch ở người) [38]. HIV là virus thuộc chi Lentivirus và thuộc họ Retrovirus có khả năng gây hội chứng suy giảm miễn dịch mắc phải (Acquired Immunodeficiency Syndrome – AIDS), một tình trạng làm hệ miễn dịch của con người bị suy giảm cấp tiến, tạo điều kiện cho những bệnh nhiễm trùng và ung thư phát triển mạnh làm đe dọa đến mạng sống của người bị nhiễm. HIV lây nhiễm vào các tế bào quan trọng trong hệ thống miễn dịch của con người như tế bào T – CD4 (tế bào bạch cầu đóng vai trò quan trọng trong hệ thống miễn dịch chống nhiễm trùng), đại thực bào và tế bào tua. Nhiễm HIV làm giảm mạnh số lượng tế bào T – CD4 thông qua ba cơ chế chính: đầu tiên, virus trực tiếp giết chết các tế bào mà chúng nhiễm vào, sau đó làm tăng tỷ lệ chết rụng tế bào ở những tế bào bị nhiễm bệnh, bước ba là các tế bào T – CD8 (tế bào bạch cầu chuyên tìm kiếm và tiêu diệt các tế bào nhiễm bệnh trong cơ thể) giết chết những tế bào T – CD4 bị nhiễm bệnh. Khi số lượng các tế bào T – CD4 giảm xuống dưới một mức giới hạn nào đó, sự miễn dịch qua trung gian tế bào bị vô hiệu và cơ thể dần dần yếu đi tạo điều kiện cho các cơ hội nhiễm trùng, cơ thể sẽ bị các mầm bệnh tấn công sinh ra nhiều chứng bệnh nguy hiểm dẫn đến cái chết. Hai loại HIV đã được định rõ đặc điểm là HIV – 1 và HIV – 2. HIV – 1 là loại virus ban đầu được phát hiện và đặt tên là LAV và HTLV – III. HIV – 1 độc hơn HIV – 2, và là nguyên nhân của phần lớn các ca nhiễm HIV trên toàn cầu. HIV – 2 có khả năng lây nhiễm thấp hơn HIV – 1 và chủ yếu trú trú gây bệnh tại Tây Phi. HIV được bao gồm hai sợi RNA, mười năm loại virus protein và một số loại protein từ tế bào vật chủ bị nhiễm. Với sự phát triển của cấu trúc mới cung cấp các huy vọng phát triển một loại vaccine phòng chống HIV. Hình 1.4 là một trong những protein gây ra bệnh HIV.



Hình 1.4 – Một trong những cấu trúc của protein HIV – 1 (pdb 3h47)

Trên đây là những ứng dụng trong rất nhiều ứng dụng của việc xác định chức năng của protein. Qua hai ứng dụng trên, chúng ta cũng có thể thấy được tầm quan trọng của việc tìm hiểu chức năng của protein.

Hai hệ thống phân loại protein phổ biến hiện nay là hệ thống phân loại SCOP (Structural Classification Of Protein) và hệ thống phân loại CATH (Class – Architecture – Topology – Homologous superfamily) để thực hiện khai thác và dự đoán cấu trúc protein. Hệ thống phân loại SCOP [22], [23] nhằm mô tả chi tiết và toàn diện về mối quan hệ cấu trúc và tiến hóa của tất cả các protein có cấu trúc được biết đến và được xây dựng bởi các giáo sư Tim J. P. Hubbard, Alexey G. Murzin, Steven E. Brenner và Cyrus Chothia đến từ đại học Cambridge. Hệ thống phân loại CATH [8] là hệ thống phân loại theo cấu trúc domain và được xây dựng bởi Christine Orengo, Janet Thornton và các cộng sự tại University College London. Khóa luận sử dụng hai cây phân loại này nhằm để có kết quả so sánh với nhau một cách khách quan.



Hình 1.5 – Các cấp độ của hai hệ thống phân loại CATH và SCOP

Dựa vào sự phân loại SCOP và CATH, hệ thống tiên hành gán nhãn cho từng protein. Và nhóm các protein lại nếu chúng có cùng nhãn. Do đó, hệ thống sẽ phân loại các cấu trúc protein mới dựa trên những cấu trúc protein có sẵn này.

Để dự đoán được cấu trúc bậc cao của protein bằng phương pháp tin học, hiện đã có nhiều mô hình được đưa ra. Jpred [34] là một chương trình dự đoán cấu trúc bậc hai thuộc trường đại học Dundee (Mỹ) công bố và đưa vào sử dụng từ năm 2000 cho tới nay. Jpred là một trong những chương trình dự đoán có kết quả hàng đầu với độ chính xác khoảng 76.4% bằng thuật toán Jnet với dữ liệu kiểm thử trên 480 protein. Gần đây, Jpred đã cải thiện thuật toán Jnet làm tăng độ chính xác tăng lên từ 81.5% tới 88.9%. CPHmodels [37] là mô hình dự đoán cấu trúc bậc ba của protein dựa trên độ tương đồng thuộc trường đại học kỹ thuật Denmark (Đan Mạch), CPHmodels dự đoán cấu trúc 3D protein bằng mạng Neural với độ chính xác 74% với tập dữ liệu huấn luyện 1377 protein và tập dữ liệu kiểm thử là 690 protein.

Ngoài một số mô hình đưa ra được áp dụng và triển khai vào thực tế ở trên, thì có một số phương pháp được áp dụng trong lĩnh vực dự đoán cấu trúc bậc cao của protein được các nhà khoa học trên thế giới nghiên cứu và đưa ra. Nguyễn Thanh Tùng [2], sử dụng mô hình Markov ẩn để dự đoán cấu trúc bậc hai của protein và chọn trình tự của chuỗi polypeptide làm đặc trưng của mỗi protein. Số mẫu làm tập dữ liệu huấn luyện là 3000, số protein là tập dữ liệu kiểm thử là 231 với độ chính xác trung bình là 63.59%. Vũ Minh Thái, Lê Hoàng Hà [4] sử mô hình ProtClass của Zeyar Aung và đã đưa ra những cải tiến để giải quyết bài toán cấu trúc 3D của protein. Với tổng dữ liệu huấn luyện là 337 protein và dữ liệu kiểm thử 172 protein. Qua kiểm thử thì độ chính xác là 74%. Zeyar Aung [13], đã xây dựng mô hình ProtClass (Protein Classification) để dự đoán cấu trúc 3D của protein dựa trên sự phân loại gần gũi giữa các protein. ProtClass sử dụng cơ sở dữ liệu SCOP để gán nhãn cho protein. Với dữ liệu huấn luyện là 540 protein và 60 protein dùng để kiểm thử. Với độ chính xác thấp nhất là 80% sau nhiều lần kiểm thử.

Một số phương pháp được áp dụng trong lĩnh vực dự đoán cấu trúc bậc cao của protein như:

- Phương pháp Bayes.
- Phương pháp K – Nearest Neighbors.
- Phương pháp cây quyết định.
- Phương pháp Markov ẩn.
- Phương pháp mạng Neural.

Các phương pháp dự đoán đã được các nhà khoa học trên thế giới nghiên cứu và phát triển rất nhiều để đạt được kết quả cao. Tuy nhiên, khi áp dụng với dữ liệu sinh học, một số vấn đề về mặt sinh học không được quan tâm nhiều trong các thuật toán làm ảnh hưởng không nhỏ đến kết quả. Một phương pháp được cho là rất thành công trong khi áp dụng cho nhiều lĩnh vực khác nhau là sử dụng thuật giải học Support Vector Machine (SVM). SVM là phương pháp học có giám sát dùng

để phân lớp. Tính chất nổi trội của SVM là đồng thời cực tiêu lỗi phân lớp và cực đại khoảng cách lè giữa các lớp. Ưu thế của SVM so với các thuật giải học khác như mạng neural, cây quyết định, bayes là giải quyết rất tốt bài toán quá khớp (dữ liệu bị nhiễu). SVM được đánh giá là một trong những phương pháp có độ chính xác rất cao và được sử dụng nhiều trong phân loại và nhận dạng (chữ viết tay, phân loại văn bản...). Vì vậy khóa luận đề xuất mô hình dự đoán cấu trúc bậc cao của protein trên cơ sở phương pháp phân lớp SVM. Bên cạnh đó, khóa luận cũng cài đặt thuật giải K – Nearest Neighbors (K-NN) để có kết quả so sánh một cách khách quan về bài toán.

Khóa luận này đã sử dụng nguồn dữ liệu từ PDB nhằm cung cấp thông tin về cấu trúc protein. Sử dụng các hệ thống phân loại phổ biến hiện nay như SCOP và CATH để phục vụ cho việc dự đoán cấu trúc protein. Khóa luận chọn bậc protein dự đoán là bậc ba bởi vì cấu trúc bậc ba của protein liên quan trực tiếp tới chức năng của protein. Do đây là vấn đề nghiên cứu còn khá mới, nên khóa luận tập trung vấn đề là tìm hiểu mối quan hệ giữa cấu trúc và chức năng từ đó dùng các kỹ thuật khai phá dữ liệu [12] để phân tích cấu trúc protein. Nghiên cứu các cấu trúc protein dựa trên các cây phân lớp có sẵn. Khi có một cấu trúc protein mới, hệ thống sẽ phân protein mới này vào một trong những lớp có sẵn.

## 1.2 Bài toán dự đoán cấu trúc bậc cao protein

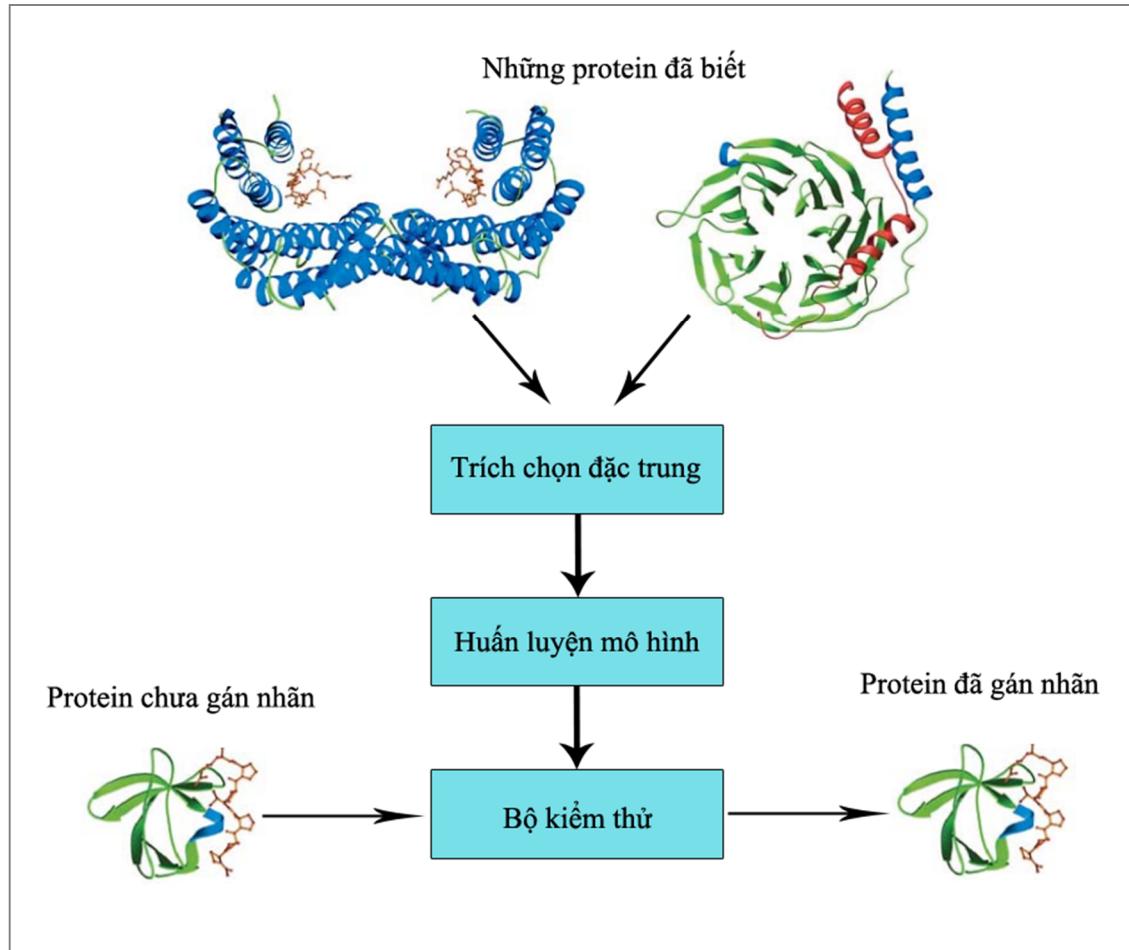
Bài toán dự đoán cấu trúc bậc cao của protein được xem là bài toán phân lớp, với các lớp chính là nhãn của protein được xác định dựa trên hai cây phân loại là SCOP và CATH.

Gọi E là tập các nhãn của mỗi nhóm protein.

Cho cấu trúc 3D của đối tượng protein  $P = \{A_c\}$ . Với  $A_c = a_1a_2a_3\dots a_n$  là chuỗi amino acid residue của chain c, c là chain của protein P.

Dự đoán cấu trúc bậc ba của protein P theo từng chain tương ứng.  $A_c = \{x_i\}$ , với  $x_i \in E$  là nhóm mà chương trình dự đoán protein P có chain c thuộc vào.

### 1.3 Mô hình hóa sơ lược bài toán



Hình 1.6 – Mô hình sơ lược về dự đoán cấu trúc protein

Ví dụ: nếu cho một protein 1a8k được lấy từ PDB để dự đoán. Protein 1a8k có bốn chain: A, B, D, E.

Input: protein 1a8k.

Output:

- Nếu sử dụng hệ thống phân loại Scop:
  - 1a8k chain A → 48724.50629 (với 48724 là Class, 48724 là Fold)
  - 1a8k chain B → 48724.50629 (với 48724 là Class, 48724 là Fold)
  - 1a8k chain D → 48724.50629 (với 48724 là Class, 48724 là Fold)
  - 1a8k chain E → 48724.50629 (với 48724 là Class, 48724 là Fold)

- Nếu sử dụng hệ thống phân loại Cath:
  - 1a8k chain A → 2.40.70 (với 2 là Class, 40 là Architecture, 70 là Topology)
  - 1a8k chain B → 2.40.70 (với 2 là Class, 40 là Architecture, 70 là Topology)
  - 1a8k chain D → 2.40.70 (với 2 là Class, 40 là Architecture, 70 là Topology)
  - 1a8k chain E → 2.40.70 (với 2 là Class, 40 là Architecture, 70 là Topology)

#### **1.4 Mục tiêu đề tài**

- Tìm hiểu cấu trúc và chức năng của protein về mặt sinh học.
- Khảo sát các mô hình dự đoán hiện tại trong và ngoài nước.
- Nghiên cứu và xây dựng mô hình dự đoán cấu trúc bậc ba của protein.
- Tìm hiểu các đặc trưng, xây dựng cơ sở dữ liệu cục bộ của đối tượng protein.
- Tìm hiểu các mô hình các cây phân loại protein trên thế giới.
- Xây dựng công cụ hỗ trợ dự đoán cấu trúc bậc ba của protein bằng hai phương pháp là SVM và K-NN.

#### **1.5 Nội dung nghiên cứu**

Đề tài tiến hành nghiên cứu các cấu trúc, các đặc trưng của protein. Phân tích các cơ sở dữ liệu từ các cơ sở dữ liệu trên thế giới và tiến hành trích rút các đặc trưng của đối tượng protein để lưu vào cơ sở dữ liệu cục bộ. Tìm hiểu các công cụ dự đoán cấu trúc protein và tiến hành cài đặt công cụ thực hiện.

Nội dung khóa luận được trình bày qua năm chương, có nội dung như sau:

- Chương 1 – Mở đầu: Nêu lên tầm quan trọng của bài toán dự đoán cấu trúc bậc cao của protein, mô tả phạm vi bài toán mà khóa luận này sẽ giải quyết, các nghiên cứu liên quan gần đây, mục tiêu của đề tài và các nội dung nghiên cứu của khóa luận.

- Chương 2 – Tổng quan về protein: Trình bày các khái niệm liên quan đến đối tượng protein, tìm hiểu chức năng, các cấu trúc của protein, nghiên cứu các cây phân loại SCOP và CATH, mô tả cách đọc dữ liệu từ file PDB, SCOP và CATH.
- Chương 3 – Xây dựng mô hình dự đoán: Trình bày lý thuyết về mô hình Support Vector Machine và k-Nearest Neighbors áp dụng cho bài toán. Cũng như cách tìm các vector đặc trưng của mỗi protein. Các công thức, cách giải quyết vấn đề cũng như các thuật toán cho việc xây dựng.
- Chương 4 – Xây dựng hệ thống chương trình: Ở phần này sẽ mô tả chi tiết cách mà khóa luận xây dựng hệ thống chương trình. Cách dự đoán một protein mới, cũng như cách sử dụng chương trình. Đưa ra một số kết quả kiểm thử và biểu đồ so sánh các kết quả với nhau.
- Chương 5 – Kết luận và kiến nghị: Phần cuối cùng sẽ trình bày về những kết quả mà khóa luận đạt được, đồng thời nêu lên hướng phát triển tiếp theo của khóa luận.

## CHƯƠNG 2

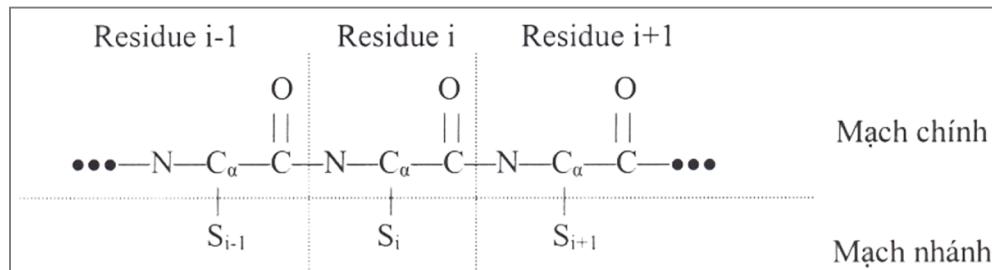
# TỔNG QUAN VỀ PROTEIN

Trong chương này, khóa luận sẽ trình bày các khái niệm liên quan tới protein, tìm hiểu các chức năng cũng như tầm quan trọng của protein về mặt sinh học. Trình bày hai loại cây phân loại phổ biến hiện nay đó là SCOP và CATH. Chương này cũng mô tả cách lấy các thông tin từ các tập tin PDB, SCOP và CATH.

## 2.1 Khái quát chung về protein

Protein được phát hiện lần đầu tiên vào năm 1745 bởi Beccari. Mới đầu được gọi là albumin (lòng trứng trắng). Mãi đến năm 1838, Mulder lần đầu tiên đưa ra thuật ngữ protein (xuất phát từ chữ Hy lạp proteos có nghĩa là “đầu tiên”, “quan trọng nhất”). Biết được tầm quan trọng và như cầu xã hội về protein, đến nay nhiều công trình nghiên cứu và sản xuất hợp chất này đã được công bố, đã đem lại nhiều ý nghĩa hết sức to lớn phục vụ cho nhân loại [1].

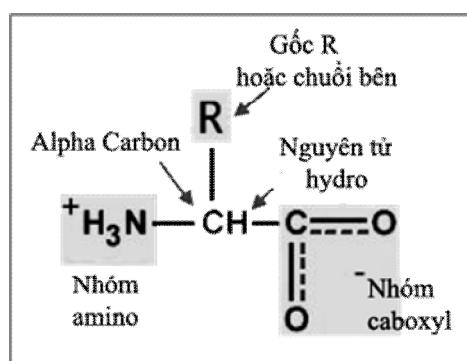
Protein [2] là những polymer sinh học được cấu tạo gồm một mạch chính (backbone hay main chain) của các đơn vị lặp lại (amino acid) với một mạch nhánh gắn vào từng đơn vị (hình 2.1).



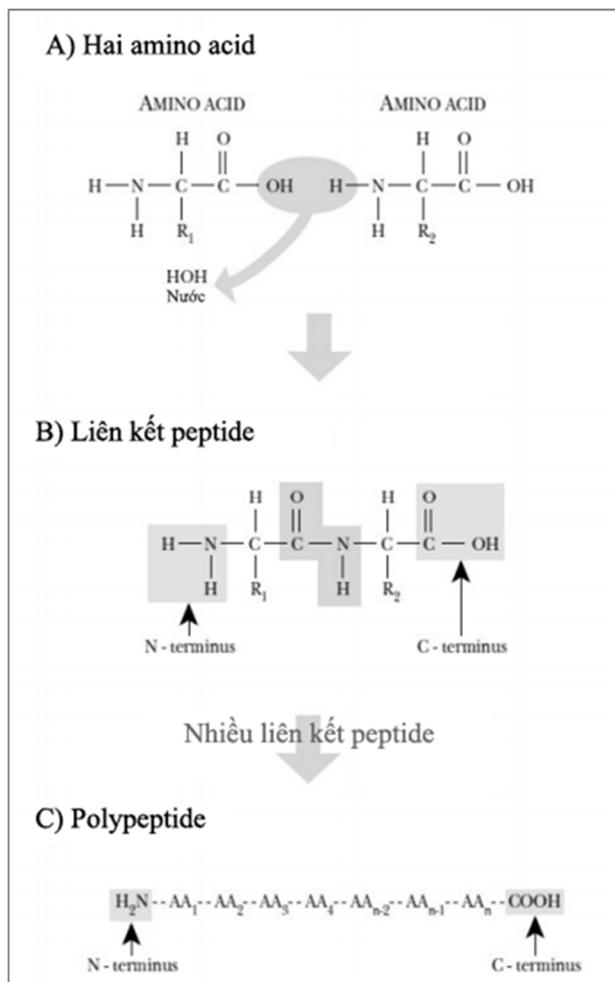
Hình 2.1 – Cấu trúc chuỗi polypeptide

Có hai mươi loại amino acid được phát hiện trong các protein của tế bào. Về cấu trúc, nói chung, mỗi amino acid gồm có một nguyên tử carbon alpha ( $C_{\alpha}$ ) trung tâm, xung quanh nó là một nhóm amino ( $-NH_2$ ), một nhóm carboxyl ( $-COOH$ ), một nguyên tử hydro ( $-H$ ) và một gốc R hay chuỗi bên đặc trưng cho từng loại amino acid (hình 2.2). Khi ở trạng thái dung dịch, các nhóm amino và carboxyl thường phân ly thành trạng thái ion, tương ứng là  $^+H_3N-$  và  $-COO^-$ . Hai amino acid nối với nhau bằng một liên kết peptide ( $-C-N-$ ) giữa nhóm carboxyl của amino acid này với

nhóm amino của amino acid kế tiếp và loại trừ một phần tử nước, cứ như thế các amino acid kết nối với nhau tạo thành một chuỗi gồm nhiều amino acid, thường được gọi là polypeptide (hình 2.3). Mỗi chuỗi polypeptide luôn có chiều xác định  $^+H_3N \rightarrow COO^-$  (do tác dụng của enzyme petydyltransferase) và được đặc trưng về số lượng, thành phần và chủ yếu là trình tự sắp xếp của các amino acid (hay còn gọi là cấu trúc sơ cấp, cấu trúc quan trọng nhất của tất cả các protein do gen quy định).



Hình 2.2 – Cấu trúc tổng quát  
của một amino acid.



Hình 2.3 – Sơ hình thành chuỗi polypeptide.

## 2.2 Chức năng sinh học của protein

Protein chiếm khoảng 60% chất hữu cơ của các cơ thể sinh vật sống [18].

Chúng tham gia chịu trách nhiệm hầu hết những phản ứng trao đổi chất và nhiều thành phần cấu trúc của tế bào. Protein có rất nhiều chức năng sinh học quan trọng khác nhau, thể hiện qua một số chức năng chính như sau:

### 2.2.1 Tạo cấu trúc

Các protein là thành phần cấu tạo cơ sở của các tế bào, bao gồm các màng tế bào, các bào quan, bộ máy di truyền của chúng. Đó cũng là các protein dạng sợi làm thành các cơ quan bộ phận trên cơ thể các động vật như: collagen làm nên xương, sụn, gân và da; keratin cấu tạo nên các lớp ngoài cùng của da và tóc, móng, sừng và lông.

### 2.2.2 Xúc tác sinh học

Hầu hết các phản ứng sinh hóa học xảy ra trong cơ thể đều do các protein đặc biệt đóng vai trò xúc tác. Những protein này được gọi là enzyme. Các enzyme đóng vai trò xúc tác cho tất cả các phản ứng hóa học trong tế bào và cơ thể đều là những protein hình cầu. Quan trọng nhất là các enzyme tham gia vào các con đường chuyển hóa và các enzyme tham gia vào các quá trình truyền thông tin di truyền trong tế bào.

### 2.2.3 Vận chuyển

Trong cơ thể động vật có xương sống, có những protein làm nhiệm vụ vận chuyển như hemoglobin, myoglobin vận chuyển O<sub>2</sub> đi khắp các mô và các cơ quan trong cơ thể hay vận chuyển ngược lại CO<sub>2</sub> về phổi để thải ra ngoài.

### 2.2.4 Vận động

Nhiều protein tham gia vào chức năng vận động của tế bào và cơ thể như actinin, myosin có vai trò vận động cơ; tubulin có vai trò vận động lông và roi của các sinh vật đơn bào.

### 2.2.5 Bảo vệ và chống đỡ

Một số protein có chức năng bảo vệ. Hệ thống các kháng thể như antibodies hay immunooglobulin là những protein đặc biệt có khả năng nhận biết và “tiêu diệt” hay “trung hòa” những chất lạ xâm nhập vào cơ thể như protein lạ, độc tố, virus, vi khuẩn. Một số protein tham gia trong quá trình đông máu như thrombin hay fibrinogen có vai trò bảo vệ cơ thể sống khỏi bị mất máu.

### 2.2.6 Truyền xung thần kinh

Một số protein có vai trò trung gian cho các phản ứng trả lời của tế bào thần kinh, đối với các kích thích đặc hiệu. Ví dụ: vai trò của sắc tố thị giác rodopsin ở võng mạc mắt.

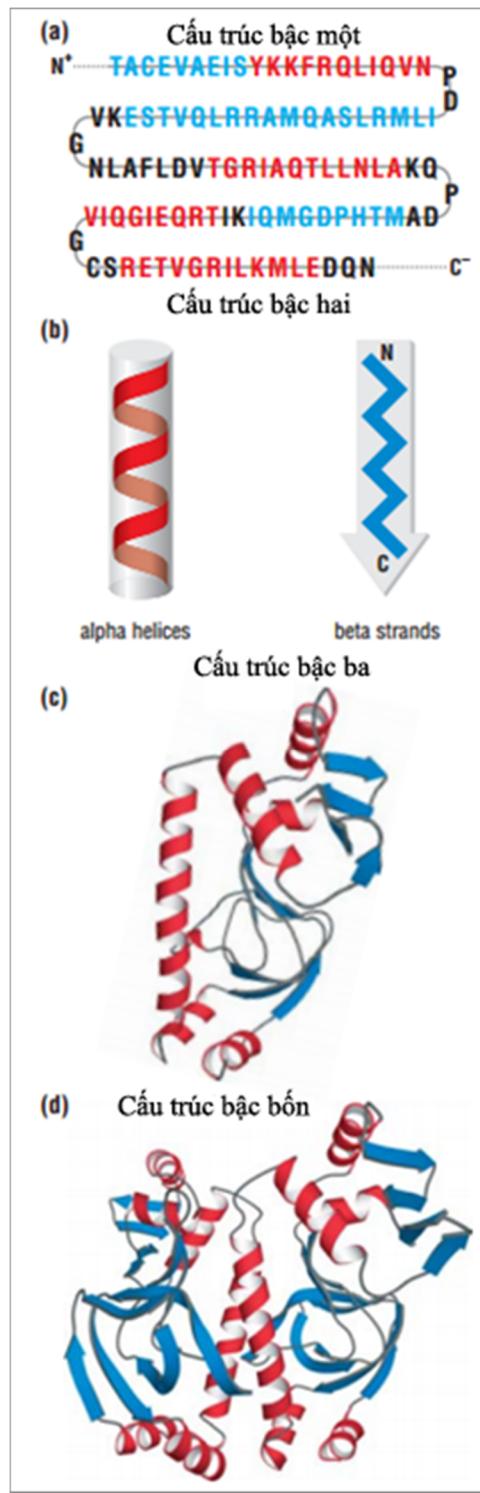
### 2.2.7 Dự trữ chất dinh dưỡng

Các protein còn là nguồn dinh dưỡng chính cung cấp năng lượng cho tế bào và cơ thể duy trì các hoạt động trao đổi chất và phát triển. Ví dụ như albumin lòng trắng trứng là nguồn cung cấp các amino acid cho phôi phát triển; casein trong sữa mẹ là nguồn cung cấp amino acid cho con; trong hạt cây có chứa nguồn protein dự trữ cần cho hạt nảy mầm.

## 2.3 Cấu trúc của protein

Cấu trúc của protein được mô tả ở bốn cấp độ, đó là:

- Cấu trúc bậc một: thực chất là trật tự sắp xếp của các amino acid trong chuỗi polypeptide.
- Cấu trúc bậc hai: là sự sắp xếp đều đặn của chuỗi polypeptide trong không gian. Chuỗi polypeptide thường không ở dạng thẳng mà chúng xoắn lại tạo nên cấu trúc xoắn alpha và nếp gấp beta nhờ liên kết hydro.
- Cấu trúc bậc ba: là một dạng không gian của cấu trúc bậc hai, làm cho phân tử protein có hình dạng gọn hơn trong không gian 3D.
- Cấu trúc bậc bốn: khi phân tử protein có nhiều chuỗi polypeptide riêng biệt phối hợp với nhau tạo nên cấu trúc bậc bốn.



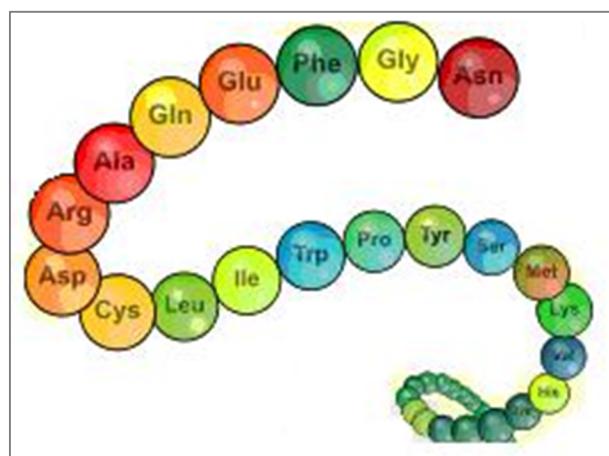
Hình 2.4 – Cấu trúc bốn cấp bậc của cấu trúc protein.  
K.Linderstrom – Lang là người đầu tiên đề xuất sự phân loại này [31].

### 2.3.1 Cấu trúc bậc một

Từ các amino acid, nhờ các liên kết peptide nối kết chúng lại với nhau tạo nên chuỗi polypeptide. Chuỗi polypeptide là cơ sở cấu trúc bậc một của protein. Tuy nhiên, không phải mọi chuỗi polypeptide đều là protein bậc một. Nhiều chuỗi polypeptide chỉ tồn tại ở dạng tự do trong tế bào mà không tạo nên phân tử protein. Những chuỗi polypeptide có trật tự amino acid xác định thì mới hình thành phân tử protein. Người ta xem cấu tạo bậc một của protein là trật tự sắp xếp của các amino acid có trong chuỗi polypeptide. Thứ tự các amino acid trong chuỗi có vai trò quan trọng vì là cơ sở cho việc hình thành cấu trúc không gian của protein và từ đó quy định đặc tính của protein.

Phân tử protein ở bậc một chưa có hoạt tính sinh học vì chưa hình thành nên các trung tâm hoạt động. Phân tử protein ở cấu trúc bậc một chỉ mang tính đặc thù về thành phần amino acid, trật tự các amino acid trong chuỗi polypeptide.

Trong tế bào protein thường tồn tại ở các bậc cấu trúc không gian. Sau khi chuỗi polypeptide của protein bậc một được tổng hợp tại ribosome, nó rời khỏi ribosome và hình thành cấu trúc không gian (bậc hai, bậc ba, bậc bốn) rồi mới di chuyển đến nơi sử dụng thực hiện chức năng của nó.



Hình 2.5 – Cấu trúc bậc một của protein

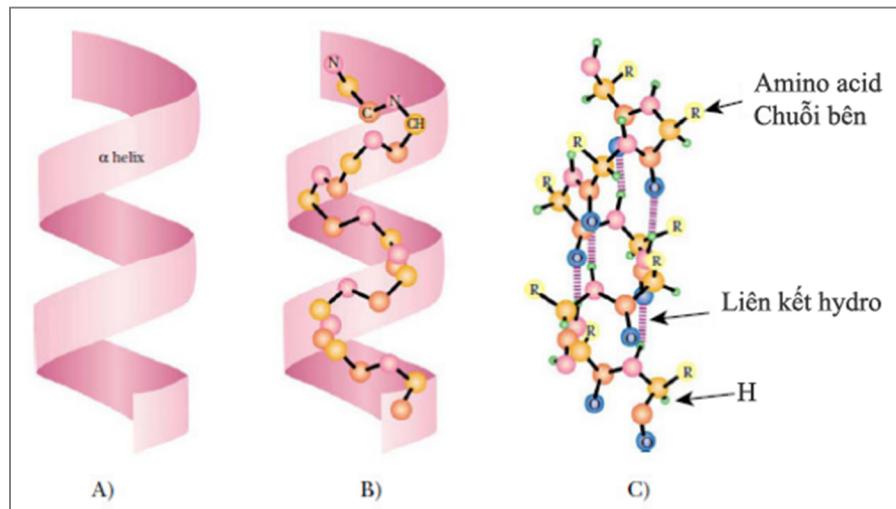
### 2.3.2 Cấu trúc bậc hai

Từ cấu trúc mạch thẳng của protein (cấu trúc bậc một), hình thành các liên kết nội phân tử, đó là các liên kết hydro làm cho chuỗi mạch thẳng cuộn xoắn lại tạo nên cấu trúc bậc hai của protein. Cấu trúc bậc hai của protein là kiểu cấu trúc không gian ba chiều.

Sở dĩ chuỗi polypeptide có thể cuộn xoắn lại được là do trong các liên kết trên chuỗi polypeptide thì liên kết peptide (C-N) là liên kết bền vững, còn các liên kết xung quanh nó ( $C_{\alpha}$ -C và  $C_{\alpha}$ -N) là các liên kết yếu, chúng có thể quay xung quanh trục của liên kết peptide.

Theo Pauling và Corey [24], [32], [33], cấu trúc bậc hai của protein có hai kiểu cấu trúc chính là xoắn alpha ( $\alpha$ -helix) và nếp gấp beta ( $\beta$ -beta sheet). Cả hai liên kết đều cho phép tạo thành lượng liên kết hydro tối đa có thể và do đó rất ổn định.

#### i. Xoắn alpha ( $\alpha$ -helix)

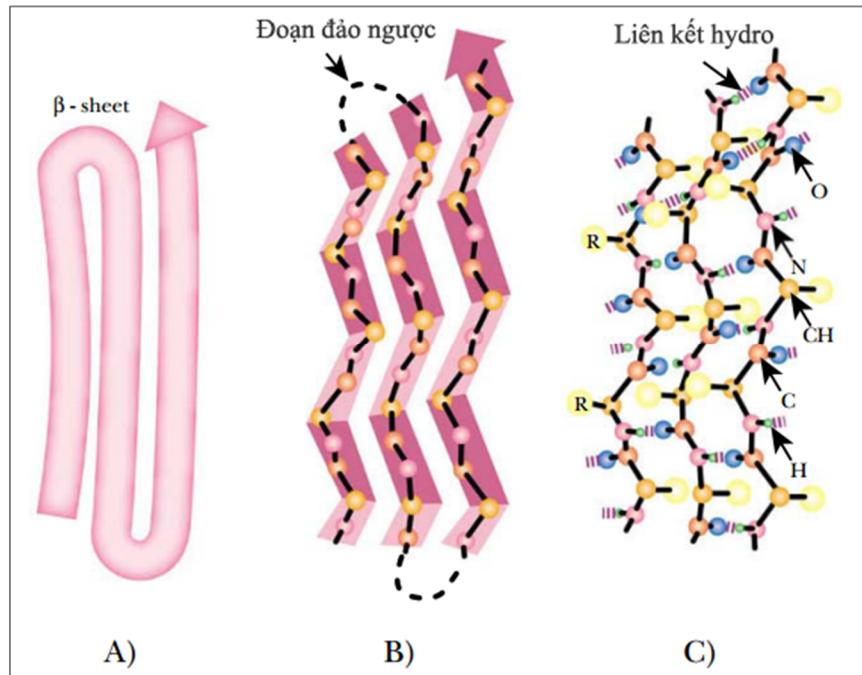


Hình 2.6 – Xoắn alpha

A) Cấu trúc chung của  $\alpha$  helix, B) Carbon backbone của chuỗi polypeptide, C) Liên kết hydro giữa hai amino acid

- Trong một chuỗi xoắn alpha (hình 2.6), một chuỗi polypeptide đơn được cuộn xoắn thành xoắn phải và liên kết hydro chạy theo chiều dọc lên và xuống, song song với trục xoắn. Trong thực tế, các liên kết hydro trong một chuỗi xoắn alpha không phải là song song với trục, nó hơi nghiêng so với trục xoắn vì mỗi vòng xoắn có 3,6 amino acid (không phải là số nguyên). Chiều dài của một vòng xoắn là 0,54 nm và chiều cao của một amino acid là 0,15 nm.
- Cấu trúc xoắn alpha được giữ vững chủ yếu nhờ liên kết hydro. Liên kết hydro chủ yếu được tạo thành giữa các nhóm  $-C=O$  của residue i với nhóm  $-NH$  của residue  $i+4$ . Chuỗi xoắn kép helix rất ổn định bởi vì tất cả các nhóm peptide ( $-NH-CO$ ) đều tham gia vào hai liên kết hydro, một ở trên và một ở dưới dọc theo trục xoắn.

## ii. Nếp gấp beta ( $\beta$ - sheet)



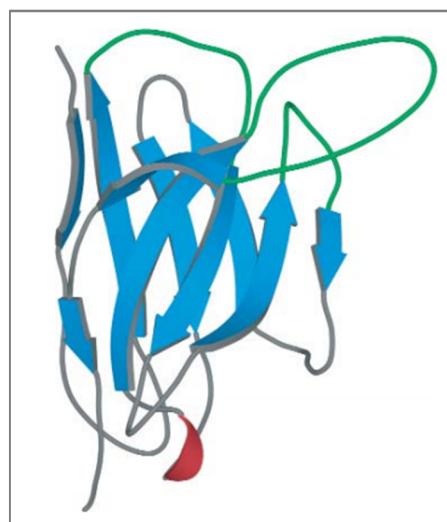
Hình 2.7 – Nếp gấp beta

- Các nếp gấp beta cũng được tổ chức với nhau bởi liên kết hydro giữa các nhóm peptide nhưng trong trường hợp này chuỗi polypeptide được gấp ngược lại trên chính nó để cho ra các cấu trúc mặt phẳng gấp theo kiểu zig-zag (hình 2.7). Cũng giống như các chuỗi xoắn alpha, các nếp gấp beta cũng rất ổn định bởi vì tất cả các nhóm peptide (ngoại trừ nhóm trên mép của dải) đều tham gia vào hai liên kết hydro. Trong các nếp gấp beta, các liên kết hydro hướng về hai bên của nhóm peptide, mỗi liên kết một bên.

Mặc dù cấu trúc của protein được cố rất nhiều bằng liên kết hydrogen, nhưng các thí nghiệm về hóa lý cho thấy rằng cấu tạo vòng xoắn ốc trong dung dịch không đủ bền và sẽ bị bung ra thành chuỗi nối. Điều này đưa đến việc đề ra cấu trúc bậc ba với nhiều loại liên kết có tác dụng ổn định hơn.

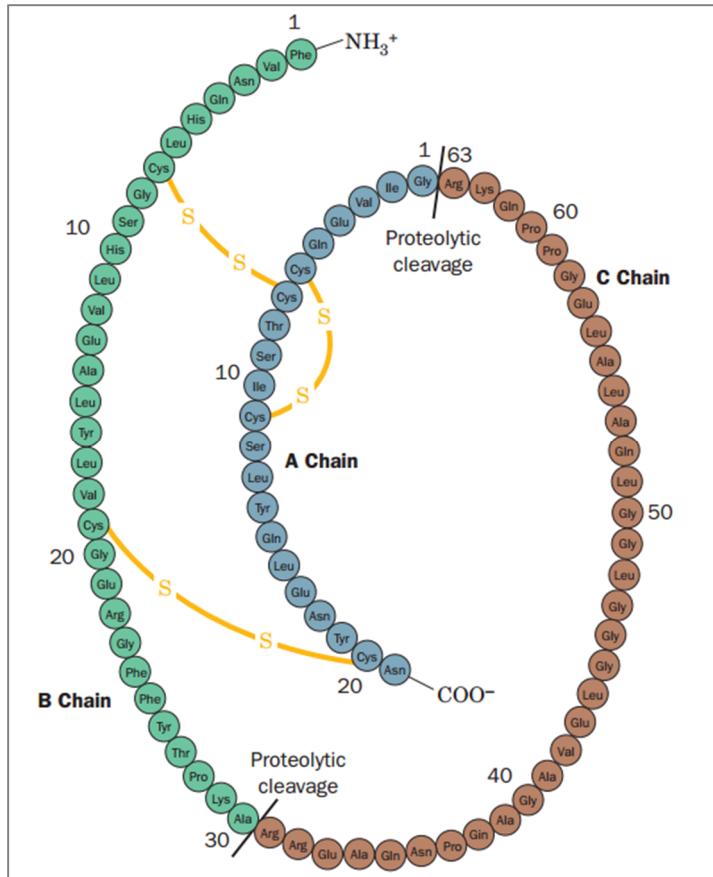
### 2.3.3 Cấu trúc bậc ba

Cấu trúc bậc ba của protein là một dạng không gian của cấu trúc bậc hai, làm cho phân tử protein có hình dạng gọn hơn trong không gian 3D. Sự thu gọn như vậy giúp cho phân tử protein ổn định trong môi trường sống.



Hình 2.8 – Cấu trúc bậc ba của protein (pdb 1ogp)

Cơ sở của cấu trúc bậc ba là liên kết disulfide (-S-S). Liên kết này được hình thành từ hai phân tử cysteine nằm xa nhau trên mạch peptide nhưng gần nhau trong cấu trúc không gian do sự cuộn lại của mạch polypeptide. Đây là liên kết đồng hóa trị nên rất bền vững.



Hình 2.9 – Liên kết disulfide trong protein

Ngoài liên kết disulfide, cấu trúc bậc ba của protein còn được ổn định (bền vững) nhờ một số liên kết khác như liên kết ion, liên kết hydro và lực van der Waals.

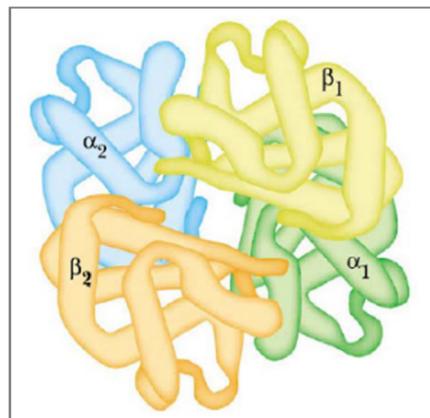
- Liên kết hydro: có thể hình thành giữa các nhóm R của hai amino acid gần nhau. Những amino acid với nhóm hydroxyl (-OH), nhóm amino hoặc amide (-NH<sub>2</sub>).

- Liên kết ion ( $\text{-NH}_3^+ \text{-OOC-}$ ): có thể được tạo thành giữa nhóm R của amino acid bazơ và amino acid axít.
- Lực Van der Waals: là lực rất yếu và giảm theo khoảng cách. Chúng chỉ quan trọng với vùng lớn về cấu hình.

Do cấu trúc bậc ba mà các protein có được hình thù đặc trưng và phù hợp với chức năng của chúng. Ở các protein chức năng như enzyme, các kháng thể... thông qua cấu trúc bậc ba mà hình thành được các trung tâm hoạt động là nơi thực hiện các chức năng của protein.

#### 2.3.4 Cấu trúc bậc bốn

Nhiều protein bao gồm nhiều chuỗi polypeptide riêng biệt. Điều này đặc biệt đúng với các protein có tổng khối lượng lớn hơn 50000 Dalton (1dvc = 1 dalton) (khoảng 400 amino acid) [Mặc dù, thỉnh thoảng tìm thấy chuỗi polypeptide có 1000 amino acid hoặc nhiều hơn thế, nhưng chúng tương đối hiếm]. Việc lắp ráp nhiều tiểu đơn vị lại với nhau tạo nên cấu trúc bậc bốn.



Hình 2.10 – Cấu trúc bậc bốn của protein

### 2.4 Cây phân loại protein

Từ khi ngân hàng dữ liệu protein (PDB-Protein Data Bank) được tạo ra phục vụ cho việc nghiên cứu và phát triển cho tới nay, một vài hệ thống phân loại cho ngân hàng dữ liệu protein trên cũng phát triển theo, đặc biệt có hai hệ thống phân loại phổ biến đó là SCOP và CATH.

#### 2.4.1 Cây phân loại SCOP (Structural Classification of Protein)

Cây phân loại SCOP [22], [23] được xây dựng bởi các giáo sư Tim J. P. Hubbar, Alexey G. Murzin, Steven E. Brenner và Cyrus Chothia thuộc đại học Cambridge. Việc phân loại cấu trúc protein trong SCOP đã được phân loại bằng tay bằng cách kiểm tra hình ảnh và so sánh các cấu trúc.

Hệ thống cây phân loại SCOP (phân loại cấu trúc protein: <http://scop.mrc-lmb.cam.ac.uk/scop/>) nhằm mục đích cung cấp mô tả một cách chi tiết và toàn diện về mối quan hệ cấu trúc và tiến hóa giữa tất cả các protein có cấu trúc được biết đến. Ưu điểm của hệ thống phân lớp này là nó đem lại kết quả có tính chính xác và tính ứng dụng rất cao.

Các cấp độ phân loại SCOP được mô tả như sau:

- Family

Protein được nhóm lại với nhau thành Family, dựa trên cơ sở là chúng xuất phát từ cùng một tổ tiên hoặc cấu trúc và chức năng của chúng tương đương nhau (với trình tự chuỗi có thể tương đương từ mức 15%).

- Superfamily

Những protein có độ tương đồng về cấu trúc chuỗi tương đối thấp, nhưng các tính năng cấu trúc và chức năng cho thấy rằng một nguồn gốc tiến hóa là có thể xảy ra thì được đặt trong Superfamily.

- Fold

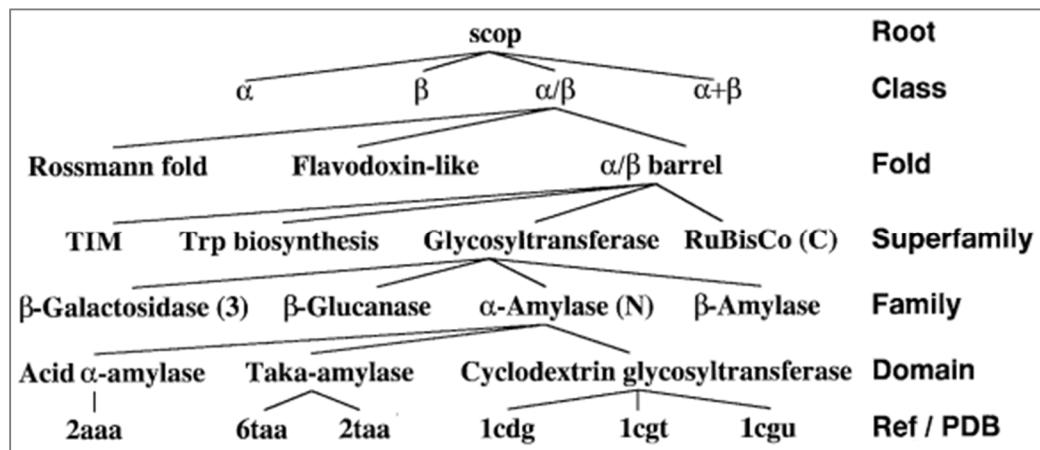
Family và Superfamily được định nghĩa là có một Fold nếu như những protein đó có cùng cấu trúc bậc hai trong cùng một cấu trúc không gian.

- Class

Các Fold khác nhau được nhóm vào một Class. Hầu hết các Fold thuộc một trong năm Class sau:

1. All –  $\alpha$ : là những cấu trúc cơ bản được hình thành bởi các alpha-helices.
2. All –  $\beta$ : là những cấu trúc cơ bản được hình thành bởi các beta-sheets.
3.  $\alpha/\beta$ : chủ yếu là các alpha – helices và beta – strains đan xen vào nhau, tạo ra các cặp beta-sheets song song với nhau.
4.  $\alpha + \beta$ : chủ yếu là các alpha – helices và beta – strains cùng hiện diện, nhưng thường chúng không đan xen vào nhau, tạo nên các cặp beta – sheets ngược với nhau.
5. Multi – domain: các protein có nhiều domain và khác cấu trúc với nhau

Các lớp khác được chỉ định vào một trong các Class sau: peptides, small proteins, theoretical models, nucleic acids và carbohydrates.



Hình 2.11 – Cấu trúc phân lớp theo kiến trúc phân loại SCOP

Phiên bản hiện tại mới nhất của SCOP là 1.75 (truy cập vào ngày 30/07/2012). Chứa đựng 38221 PDB và 110800 Domain.

Bảng 2.1 – Bảng thống kê hệ thống phân loại SCOP version 1.75

| Class                                     | Fold | Superfamily | Family |
|---|------|-------------|--------|
| All alpha protein                         | 284  | 507         | 871    |
| All beta protein                          | 174  | 354         | 742    |
| Alpha and beta protein ( $\alpha/\beta$ ) | 147  | 244         | 803    |
| Alpha and beta protein ( $\alpha+\beta$ ) | 376  | 552         | 1055   |
| Multi – domain protein                    | 66   | 66          | 89     |
| Membrane and cell surface proteins        | 58   | 110         | 123    |
| Small proteins                            | 90   | 129         | 219    |
| Tổng số lượng protein                     | 1195 | 1962        | 3902   |

#### 2.4.2 Cây phân loại CATH (Class – Architecture – Topology – Homologous superfamily)

Phân loại cấu trúc protein CATH [8] là hệ thống phân loại bán tự động (hơn 90% protein trong CATH là phân cấp tự động), cây phân loại protein được phát triển bởi Christine Orengo, Janet Thornton và các cộng sự tại University College London.

Hệ thống cây phân loại CATH (<http://www.cathdb.info>) là hệ thống phân loại theo cấu trúc domain khác với cách tiếp cận của SCOP. Cây phân loại CATH phân cấp theo bốn cấp độ chính, đó là: Class (C), Architecture (A), Topology (T), Homologous Superfamily (H).

Phương thức tự động trong CATH: sử dụng chỉ số SSAP (Sequential Structure Alignment Program, phương pháp tự động so sánh cấu trúc protein bằng trình tự chuỗi) và so sánh chuỗi. Khi mức độ tương đồng của trình tự chuỗi  $\geq 35\%$  và chỉ số SSAP  $\geq 80\%$  thì domain của protein đó được phân vào cùng cấp với domain đang được xét. Ngược lại thì domain của protein đó được phân loại bằng tay.

Các cấp độ của CATH được mô tả như sau:

- Homologous Superfamily (cấp độ H)

Tại cấp độ H (tương ứng với cấp độ Superfamily của SCOP) là cấp độ nhóm các domain protein theo phương pháp tự động bằng phương pháp SSAP. Phương pháp này được thực hiện như sau:

- Mức tương đồng trình tự chuỗi giữa các domains  $\geq 35\%$ , tỉ lệ chòng chéo  $\geq 60\%$  giữa các cấu trúc lớn và cấu trúc nhỏ.
- Chỉ số SSAP  $\geq 80\%$ , mức độ tương đồng giữa các domains  $\geq 20\%$ , tỉ lệ chòng chéo  $\geq 60\%$ .
- Chỉ số SSAP  $\geq 80\%$ , tỉ lệ chòng chéo  $\geq 60\%$ , và các domain đều có chức năng tương tự nhau.

- Topology (cấp độ T)

Những cấu trúc được nhóm vào cùng một cấp độ T (tương ứng với cấp độ Fold của SCOP) tùy thuộc vào hình dạng tổng thể và kết nối của cấu trúc bậc hai. Điều này được thực hiện bằng cách sử dụng thuật toán SSAP.

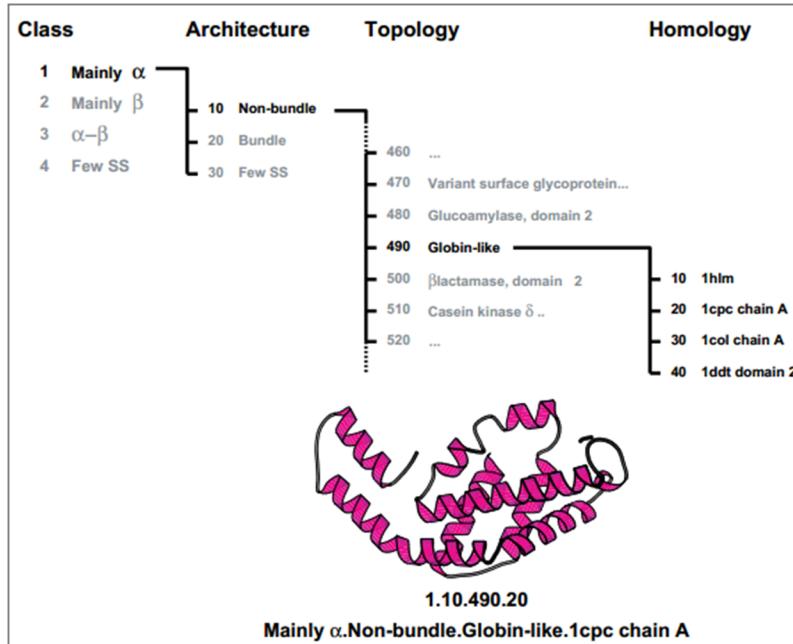
- Architecture (cấp độ A)

Đây là cấp độ mô tả hình dạng tổng thể cấu trúc domain được xác định bởi các định hướng của cấu trúc bậc hai nhưng bỏ qua các kết nối giữa các cấu trúc bậc hai.

- Class (cấp độ C)

Cấp độ C tương ứng với cấp độ Class trong SCOP. Cấp độ C có 3 loại chính được:

- Mainly – alpha
- Mainly – beta
- Alpha – beta (bao gồm các cấu trúc alpha/beta và alpha + beta)



Hình 2.12 – Cấu trúc phân lớp theo kiến trúc phân loại CATH

Phiên bản hiện tại mới nhất của CATH là 3.4 (truy cập vào ngày 30/07/2012).

Chứa đựng 104,238 chain protein.

Bảng 2.2 – Bảng thống kê hệ thống phân loại CATH version 3.4

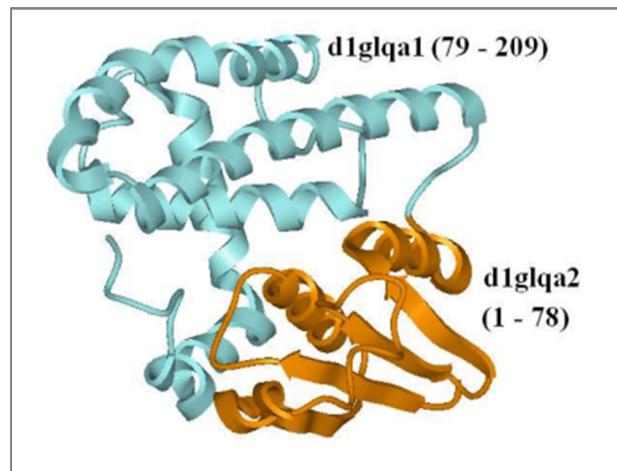
| Class | Architecture | Topology | Homologous Superfamily | Domain |
|-------|--------------|----------|------------------------|--------|
| 1     | 5            | 376      | 839                    | 32396  |
| 2     | 20           | 228      | 514                    | 39140  |
| 3     | 14           | 577      | 1082                   | 79038  |
| 4     | 1            | 101      | 114                    | 2346   |
| Total | 40           | 1282     | 2549                   | 152920 |

## 2.5 Định nghĩa cấu trúc domain trong protein

Khái niệm về domain lần đầu tiên được đề xuất vào năm 1973 bởi Wetlaufer [16]. Wetlaufer định nghĩa domain như là các đơn vị ổn định của cấu trúc protein có thể gấp một cách tự động. Các domain có thể có từ 25 amino acid cho tới 500 amino acid trong một protein.

Cấu trúc domain có thể được hiểu như là những bộ phận, những khu vực trong một phân tử protein được cuộn gấp trong không gian 3D giống như một phân tử protein nhỏ hoàn chỉnh và thường là những nơi thực hiện chức năng liên kết, chức năng lắp ráp của phân tử protein trong hoạt động chức năng của nó.

Sự hình thành các domain trong phân tử protein tạo ra khả năng tương tác linh hoạt giữa các đại phân tử, khả năng cơ động, dịch chuyển tương ứng giữa những bộ phận trong quá trình thực hiện chức năng sinh học. Ở những protein có nguồn gốc khác nhau, nhưng có chức năng tương tự thì các domain có cấu trúc tương đối giống nhau.



Hình 2.13 – Hai domain trong protein 1glqA dựa vào cây phân loại SCOP

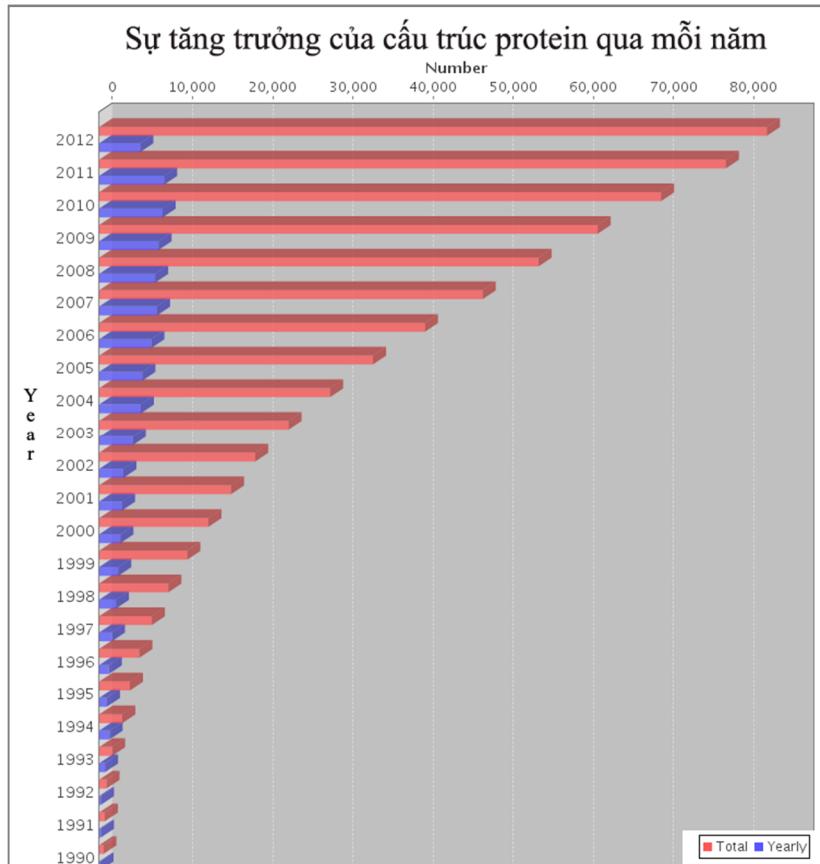
Hình 2.13 là một ví dụ về cấu trúc domain của protein 1glqA dựa vào cây phân loại SCOP. Protein 1glqA có hai domain, domain d1glqa1 có residue từ 79 đến 209 và domain d1gla2 có residue từ 1 đến 78.

## 2.6 Tài nguyên thông tin cấu trúc protein

### 2.6.1 Thông tin cấu trúc 3D của protein

Thông tin về cấu trúc 3D của các đại phân tử sinh học có thể được tìm thấy trong dữ liệu của PDB (Protein Data Bank – <http://www.pdb.org>). Hầu hết các cấu trúc dữ liệu trong cơ sở dữ liệu PDB đã thu được chủ yếu bởi một trong hai phương

pháp: x-quan tinh thê (trên 80%), giải pháp cộng hưởng từ hạt nhân (NMR) (khoảng 16%) và còn lại là các phương pháp khác.



Hình 2.14 – Sự tăng trưởng của cơ sở dữ liệu PDB trong những năm qua (PDB Static, 24/07/ 2012)

Vào ngày 24 tháng 07 năm 2012, kho lưu trữ PDB đã lưu giữ 83,266 cấu trúc protein. Tổng dung lượng chưa nén là hơn 750 GB. Số lượng của cơ sở dữ liệu PDB tăng nhanh trong những năm gần đây do có sự tiến bộ của khoa học và công nghệ. Thông tin cấu trúc của protein gồm có các tọa độ 3D của các atom của một hay nhiều phân tử mà một protein chứa chúng. Các tọa độ nguyên tử này cũng được gọi là cấu trúc 3D hay còn gọi là cấu trúc bậc ba. Cấu trúc bậc ba của một protein gắn chặt với chức năng của nó. Vì vậy, việc hiểu rõ cấu trúc bậc ba thường giúp hiểu rõ chức năng bên trong của protein.

### 2.6.2 Đọc thông tin từ PDB

Ở đây, khóa luận chỉ trình bày những gì mà có liên quan tới, các phần còn lại có thể tham khảo tại Protein Data Bank Contents Guide: Atome Coordinate Entry Format Description [20]. Phiên bản hiện tại là 3.30 (ngày cập nhật 17/01/2011).

Mỗi file PDB chứa đựng toàn bộ thông tin có liên quan tới cấu trúc protein. Các thông tin như cấu trúc 3D, chuỗi amino acid, yếu tố nhiệt độ của từng atom. Như đã trình bày ở phần 2.1, mỗi amino acid có một carbon alpha là carbon trung tâm. Do đó, chỉ cần lấy gốc atom có chứa  $C_\alpha$  làm đặc trưng cho mỗi amino acid và không quan tâm tới những gốc atom không chứa  $C_\alpha$  của cùng một amino acid.

Bảng 2.3 – Mô tả chi tiết các cột atom của file protein PDB

| Cột     | Kiểu dữ liệu | Tên thuộc tính | Mô tả                       |
|---------|--------------|----------------|-----------------------------|
| 1 – 6   | Record name  | “ATOM”         | “ATOM”                      |
| 7 – 11  | Intenger     | Serial         | Số thứ tự của Atom          |
| 13 – 16 | Atom         | Name           | Tên Atom                    |
| 17      | Character    | altLoc         | Thay thế vị trí chỉ định    |
| 18 – 20 | Residue name | resName        | Tên residue                 |
| 22      | Character    | chainId        | Mã chain                    |
| 23 – 36 | Intenger     | resSeq         | Số thứ tự residue           |
| 27      | AChar        | iCode          | Mã thêm vào residues        |
| 31 – 38 | Real(8.3)    | X              | Tọa độ X trong Angstroms    |
| 39 – 46 | Real(8.3)    | Y              | Tọa độ Y trong Angstroms    |
| 47 – 54 | Real(8.3)    | Z              | Tọa độ Z trong Angstroms.   |
| 55 – 60 | Real(6.2)    | Occupancy      | Phòng                       |
| 61 – 66 | Real(6.2)    | tempFactor     | Yếu tố nhiệt độ             |
| 77 – 78 | LString(2)   | Element        | Biểu tượng số nguyên tố     |
| 79 – 80 | LString(2)   | Charge         | Phí của nguyên tử trên Atom |

Các atom trình bày tọa độ 3D của các amino acid và nucleotides. Bên dưới là một ví dụ về atom của protein 1glq.

|      | 1        | 2     | 3       | 4   | 5      | 6      | 7       | 8          |   |
|------|----------|-------|---------|-----|--------|--------|---------|------------|---|
| ATOM | 1        | N     | PRO A   | 1   | 71.393 | -3.633 | -4.205  | 1.00 19.20 | N |
| ATOM | 2        | CA    | PRO A   | 1   | 70.301 | -4.557 | -3.979  | 1.00 18.50 | C |
| ATOM | 3        | C     | PRO A   | 1   | 70.930 | -5.713 | -3.201  | 1.00 20.58 | C |
| ATOM | 4        | O     | PRO A   | 1   | 72.163 | -5.661 | -3.016  | 1.00 20.71 | O |
| ATOM | 5        | CB    | PRO A   | 1   | 69.792 | -4.952 | -5.349  | 1.00 19.36 | C |
| ATOM | 6        | CG    | PRO A   | 1   | 70.615 | -4.136 | -6.332  | 1.00 20.18 | C |
| ATOM | 7        | CD    | PRO A   | 1   | 71.068 | -2.995 | -5.461  | 1.00 20.18 | C |
| ATOM | 8        | N     | PRO A   | 2   | 70.234 | -6.726 | -2.687  | 1.00 18.71 | N |
| ATOM | 9        | CA    | PRO A   | 2   | 68.766 | -6.840 | -2.682  | 1.00 18.85 | C |
| ATOM | 10       | C     | PRO A   | 2   | 68.027 | -5.809 | -1.804  | 1.00 16.93 | C |
| ATOM | 11       | O     | PRO A   | 2   | 68.667 | -5.226 | -0.920  | 1.00 16.21 | O |
| ATOM | 12       | CB    | PRO A   | 2   | 68.566 | -8.264 | -2.261  | 1.00 19.84 | C |
| ATOM | 13       | CG    | PRO A   | 2   | 69.752 | -8.610 | -1.376  | 1.00 18.95 | C |
| ATOM | 14       | CD    | PRO A   | 2   | 70.866 | -7.878 | -2.065  | 1.00 18.55 | C |
| .    | .        | .     | .       | .   | .      | .      | .       | .          | . |
| .    | Tên Atom | Chain | Residue | X   | Y      | Z      | .       | .          | . |
| ATOM | 3295     | N     | VAL H   | 226 | 2.242  | 0.681  | -15.076 | 1.00 16.56 | N |
| ATOM | 3296     | CA    | VAL H   | 226 | 1.450  | 0.524  | -16.287 | 1.00 24.87 | C |
| ATOM | 3297     | C     | VAL H   | 226 | 1.967  | 1.427  | -17.393 | 1.00 22.37 | C |
| ATOM | 3298     | O     | VAL H   | 226 | 2.521  | 2.507  | -17.165 | 1.00 18.82 | O |
| ATOM | 3299     | CB    | VAL H   | 226 | -0.051 | 0.733  | -16.054 | 1.00 30.69 | C |
| ATOM | 3300     | CG1   | VAL H   | 226 | -0.563 | -0.183 | -14.950 | 1.00 31.06 | C |
| ATOM | 3301     | CG2   | VAL H   | 226 | -0.375 | 2.173  | -15.714 | 1.00 22.85 | C |
| ATOM | 3302     | N     | PRO H   | 227 | 1.817  | 0.993  | -18.636 | 1.00 19.95 | N |
| ATOM | 3303     | CA    | PRO H   | 227 | 2.404  | 1.790  | -19.717 | 1.00 22.24 | C |
| ATOM | 3304     | C     | PRO H   | 227 | 1.792  | 3.180  | -19.782 | 1.00 23.97 | C |
| ATOM | 3305     | O     | PRO H   | 227 | 0.609  | 3.398  | -19.513 | 1.00 28.61 | O |
| ATOM | 3306     | CB    | PRO H   | 227 | 2.040  | 1.005  | -20.977 | 1.00 28.53 | C |
| ATOM | 3307     | CG    | PRO H   | 227 | 1.624  | -0.348 | -20.520 | 1.00 27.91 | C |
| ATOM | 3308     | CD    | PRO H   | 227 | 1.064  | -0.170 | -19.137 | 1.00 28.69 | C |

Hình 2.15 – Định dạng tọa độ 3D của protein 1glq trong PDB

### 2.6.3 Đọc thông tin từ cây phân loại SCOP

Cây phân loại SCOP mô tả chi tiết và đầy đủ các protein đã biết trong file “dir.cla.scop.txt”. Tất cả các tài liệu có liên quan tới SCOP được cung cấp đầy đủ tại trang chủ của SCOP.

Ở hình 2.16 mô tả thông tin của mỗi domain protein từ cây phân loại SCOP.

Protein 1g1q chain A có hai domain d1g1qa1 và d1g1qa2. Thông tin cấu trúc phân lớp của domain d1g1qa1 là: class = 46456, fold = 47615, superfamily = 17616, family = 47617, điểm bắt đầu residue là 79 và điểm kết thúc residue là 209. Tương tự domain d1g1qa2 có class = 51349, fold = 52832, superfamily = 52833, family = 52862, điểm bắt đầu residue là 1 và điểm kết thúc residue là 78.

| #                    | dir.cla.scop.txt   |   |  |   |      |             |        |   |
|----------------------|--|---|--|---|------|-------------|--------|---|
|                      | # SCOP release 1.75 (June 2009) [File format version 1.00] |   |  |   |      |             |        |   |
|                      | # http://scop.mrc-lmb.cam.ac.uk/scop/                      |   |  |   |      |             |        |   |
|                      |  | Điểm bắt đầu domain   | Điểm kết thúc domain   | Class   | Fold | Superfamily | Family | . |
|                      |  | .   | .  | .   | .    | .           | .      | . |
| d1dlwa_ 1dlw         | A: a.1.1.1 14982   | cl=46456, cf=46457, sf=46458, fa=46459, dm=46460, sp=46461, px=14982  |  |   |      |             |        |   |
| d1uvya_ 1uvy         | A: a.1.1.1 100068  | cl=46456, cf=46457, sf=46458, fa=46459, dm=46460, sp=46461, px=100068 |  |   |      |             |        |   |
| d1dlya_ 1dly         | A: a.1.1.1 14983   | cl=46456, cf=46457, sf=46458, fa=46459, dm=46460, sp=46462, px=14983  |  |   |      |             |        |   |
| d1uvxu_ 1uvxu        | A: a.1.1.1 100067  | cl=46456, cf=46457, sf=46458, fa=46459, dm=46460, sp=46462, px=100067 |  |   |      |             |        |   |
|                      | .  | Điểm bắt đầu domain   | Điểm kết thúc domain   | .   | .    | .           | .      | . |
| d2gsra1_ 2gsr        | A: 77-207  | a.45.1.1  | 17586  | cl=46456, cf=47615, sf=47616, fa=47617, dm=81347, sp=47620, px=17586        |      |             |        |   |
| d2gsrb2_ 2gsr        | B: 77-207  | a.45.1.1  | 17587  | cl=46456, cf=47615, sf=47616, fa=47617, dm=81347, sp=47620, px=17587        |      |             |        |   |
| <b>d1g1qa1_ 1g1q</b> | <b>A: 79-209</b>   | <b>a.45.1.1</b>   | <b>17588</b>   | <b>cl=46456, cf=47615, sf=47616, fa=47617, dm=81347, sp=47621, px=17588</b> |      |             |        |   |
| d1g1qb1_ 1g1q        | B: 79-209  | a.45.1.1  | 17589  | cl=46456, cf=47615, sf=47616, fa=47617, dm=81347, sp=47621, px=17589        |      |             |        |   |
| d1g1qa1_ 1g1p        | A: 79-209  | a.45.1.1  | 17590  | cl=46456, cf=47615, sf=47616, fa=47617, dm=81347, sp=47621, px=17590        |      |             |        |   |
| Mã scop              | Mã protein   | Chain   |  | Class   | Fold | Superfamily | Family | . |
| d2gsra2_ 2gsr        | A: 1-76  | c.47.1.5  | 32880  | cl=51349, cf=52832, sf=52833, fa=52862, dm=81358, sp=52865, px=32880        |      |             |        |   |
| d2gsrb2_ 2gsr        | B: 1-76  | c.47.1.5  | 32881  | cl=51349, cf=52832, sf=52833, fa=81358, sp=52865, px=32881                  |      |             |        |   |
| <b>d1g1qa2_ 1g1q</b> | <b>A: 1-78</b>   | <b>c.47.1.5</b>   | <b>32882</b>   | <b>cl=51349, cf=52832, sf=52833, fa=52862, dm=81358, sp=52866, px=32882</b> |      |             |        |   |
| d1g1qb2_ 1g1q        | B: 1-78  | c.47.1.5  | 32883  | cl=51349, cf=52832, sf=52833, fa=52862, dm=81358, sp=52866, px=32883        |      |             |        |   |
| d1g1qa2_ 1g1p        | A: 1-78  | c.47.1.5  | 32884  | cl=51349, cf=52832, sf=52833, fa=52862, dm=81358, sp=52866, px=32884        |      |             |        |   |
| .                    | .  | .   | .  | .   | .    | .           | .      | . |
| d1tjba_ 1tjb         | A: k.44.1.1  | 107024  | cl=58788, cf=111554, sf=111555, fa=111556, dm=111557, sp=111558, px=107024 |   |      |             |        |   |
| d1tjbb_ 1tjb         | B: k.44.1.1  | 107025  | cl=58788, cf=111554, sf=111555, fa=111556, dm=111557, sp=111558, px=107025 |   |      |             |        |   |
| d1pyza_ 1pyz         | A: k.39.1.1  | 111646  | cl=58788, cf=90314, sf=90315, fa=90316, dm=90317, sp=90318, px=111646      |   |      |             |        |   |
| d1pyzb_ 1pyz         | B: k.39.1.1  | 111647  | cl=58788, cf=90314, sf=90315, fa=90316, dm=90317, sp=90318, px=111647      |   |      |             |        |   |
| d1vl3a_ 1vl3         | A: k.39.1.1  | 108716  | cl=58788, cf=90314, sf=90315, fa=90316, dm=90317, sp=90318, px=108716      |   |      |             |        |   |
| d1vl3b_ 1vl3         | B: k.39.1.1  | 108717  | cl=58788, cf=90314, sf=90315, fa=90316, dm=90317, sp=90318, px=108717      |   |      |             |        |   |
| d1pbza_ 1pbz         | A: k.39.1.1  | 94424   | cl=58788, cf=90314, sf=90315, fa=90316, dm=103800, sp=103801, px=94424     |   |      |             |        |   |
| d1pbzb_ 1pbz         | B: k.39.1.1  | 94425   | cl=58788, cf=90314, sf=90315, fa=90316, dm=103800, sp=103801, px=94425     |   |      |             |        |   |

Hình 2.16 – Thông tin cây phân loại SCOP cho hai domain của protein 1g1qA

#### 2.6.4 Đọc thông tin từ cây phân loại CATH

Thông tin hướng dẫn từ “CATH Domain Description File (CDDF) Format 2.0”.

Tất cả các tài liệu liên quan tới CATH được cung cấp đầy đủ tại trang chủ CATH.

Bảng 2.4 – Định dạng thông tin từ cây phân loại CATH

| STT | Tên cột   | Mô tả  |
|-----|-----------|--|
| 1   | FORMAT    | Mô tả phiên bản định dạng và là hàng đầu tiên            |
| 2   | DOMAIN    | Mã domain trong CATH – có bảy ký tự                      |
| 3   | VERSION   | Phiên bản CATH   |
| 4   | VERDATE   | Ngày phát hành phiên bản CATH                            |
| 5   | NAME      | Mô tả tên PDB  |
| 6   | SOURCE    | Mô tả cấp độ   |
| 7   | CATHCODE  | Mã CATH  |
| 8   | CLASS     | Mô tả cấp độ class                                       |
| 9   | ARCH      | Mô tả cấp độ architecture                                |
| 10  | TOPOL     | Mô tả cấp độ topology                                    |
| 11  | HOMOL     | Mô tả cấp độ homologous superfamily                      |
| 12  | DLENGTH   | Chiều dài của chuỗi domain                               |
| 13  | DSEQH     | Tiêu đề chuỗi domain trong định dạng FASTA               |
| 14  | DSEQS     | Chuỗi domain trong định dạng FAST                        |
| 15  | NSEGMENTS | Số phân đoạn của domain                                  |
| 16  | SEGMENT   | Chuỗi phân đoạn  |
| 17  | SRANGE    | Điểm bắt đầu và kết thúc của PDB residue trong phân đoạn |
| 18  | SLENGTH   | Chiều dài của chuỗi phân đoạn                            |
| 19  | SSEQH     | Tiêu đề chuỗi phân đoạn trong định dạng FASTA            |
| 20  | SSEQS     | Chuỗi phân đoạn trong định dạng FASTA                    |
| 21  | ENDSEG    | Dấu hiệu kết thúc  |

```

#-----#
# FILE NAME:      CathDomainDescriptionFile.v3.4.0
# FILE DATE:      21.11.2010
# .
.
.
FORMAT    CDDF1.0
DOMAIN    1g1qA02 ← Mã Cath
VERSION   3.4.0
VERDATE   21-Nov-2010
NAME      P-selectin. Chain: a, b, c, d. Fragment: lectin/egf domains. Synonym:
NAME      granule membrane protein 140, gmp-140, padgem, cd62p, leukocyte-endoth
NAME      elial cell adhesion molecule 3, lecam3. Engineered: yes
SOURCE    Homo sapiens. Human. Organism_taxid: 9606. Expressed in: cricetus g
SOURCE    riseus.
CATHCODE  2.10.25.10
CLASS     Mainly Beta
ARCH     Ribbon → Homology
TOPOL    Laminin → Topology
HOMOL    Laminin → Architecture
DLENGTH   38 → Class
DSEQH    > pdb|1g1qA02
DSEQS    ASCQDMSCSKQGECLETIGNYTCSCYPGFYGPECEYVR
NSEGMENTS 1 → Điểm bắt đầu domain
SEGMENT   1g1qA02:1:1 → Điểm bắt đầu domain
SRANGE    START=120 STOP=157 → Điểm kết thúc domain
SLENGTH   38 → Điểm kết thúc domain
SSEQH    > pdb|1g1qA02:1:1 → Điểm kết thúc domain
SSEQS    ASCQDMSCSKQGECLETIGNYTCSCYPGFYGPECEYVR
ENDSEG
.
.
.
FORMAT    CDDF1.0
DOMAIN    1g1qA01
VERSION   3.4.0
VERDATE   21-Nov-2010
NAME      P-selectin. Chain: a, b, c, d. Fragment: lectin/egf domains. Synonym:
NAME      granule membrane protein 140, gmp-140, padgem, cd62p, leukocyte-endoth
NAME      elial cell adhesion molecule 3, lecam3. Engineered: yes
SOURCE    Homo sapiens. Human. Organism_taxid: 9606. Expressed in: cricetus g
SOURCE    riseus.
CATHCODE  3.10.100.10
CLASS     Alpha Beta
ARCH     Roll
TOPOL    Mannose-Binding Protein A; Chain A
HOMOL    Mannose-Binding Protein A, subunit A
DLENGTH   119
DSEQH    > pdb|1g1qA01
DSEQS    WTYHYSTKAYSWNISRKYCQNRYTDLVAIQNKNEIDYLNKVLPPYSSYYWIGIRKNNKTWTWVGTKKALT
DSEQS    NEAEWNADNEPNKRNNEDCVEIYIKSPSAPGKWDEHCLKKKHALCYT
NSEGMENTS 1
SEGMENT   1g1qA01:1:1
SRANGE    START=1 STOP=119
SLENGTH   119
SSEQH    > pdb|1g1qA01:1:1
SSEQS    WTYHYSTKAYSWNISRKYCQNRYTDLVAIQNKNEIDYLNKVLPPYSSYYWIGIRKNNKTWTWVGTKKALT
SSEQS    NEAEWNADNEPNKRNNEDCVEIYIKSPSAPGKWDEHCLKKKHALCYT
ENDSEG
//
```

Hình 2.17 – Thông tin cây phân loại CATH cho hai domain của protein 1g1qA.

Hình 2.17 mô tả thông tin của mỗi domain protein từ cây phân loại CATH. Protein 1g1q chain A có hai domain 1g1qA01 và 1g1qA02. Thông tin cấu trúc phân lớp của domain 1g1qA01 là: class = 3, architecture = 10, topology = 100, homology = 10, điểm bắt đầu residue là 1 và điểm kết thúc residue là 119. Tương tự domain 1g1qA02 có class = 2, architecture = 10, topology = 25, homology = 10, điểm bắt đầu residue là 120 và điểm kết thúc residue là 157.

## 2.7 Tổng kết chương 2

Các thông tin protein được lấy từ PDB dùng để phân tích cấu trúc không gian của protein. Các thông tin này chủ yếu là các tọa độ của từng atom, loại atom hay các thông tin về mã protein đó trong pdb. Mỗi protein có thể có một hay nhiều chain. Các chain trong một protein có thể thuộc cùng một lớp hoặc khác lớp. Do đó, việc dự đoán cấu trúc protein mới sẽ tiến hành theo từng chain.

Khóa luận chọn cấu trúc bậc ba của protein là cấp bậc dự đoán vì cấu trúc bậc ba của protein liên quan trực tiếp tới chức năng sinh học của chúng. Và chọn cấp độ Fold trong cây phân loại SCOP và cấp độ Topology trong cây phân loại CATH để dự đoán. Hai cấp độ này là tương đương với nhau và được dùng để gán nhãn cho từng protein. Ở cấp độ phân lớp này, nếu dự đoán là chính xác thì đã giúp cho các nhà sinh học rất nhiều trong việc phân tích chức năng của protein.

Trong chương này, khóa luận đã giới thiệu về những khái niệm, chức năng và các cấp bậc của protein cũng như tìm hiểu hai cây phân loại phổ biến hiện nay đó là SCOP và CATH. Qua chương tiếp theo, khóa luận sẽ trình bày chi tiết cách xây dựng mô hình dự đoán, cách so sánh cấu trúc protein dựa trên độ tương đồng cũng như cách tìm đặc trưng của mỗi protein.

## **CHƯƠNG 3**

# **XÂY DỰNG MÔ HÌNH DỰ ĐOÁN**

Việc khảo sát các phương pháp máy học được áp dụng thành công cho nhiều bài toán khác nhau (các phương pháp như Bayes, cây quyết định, K – Nearest Neighbors, Markov ẩn, mạng Neural, Support Vector Machine) cho thấy có khá nhiều phương pháp máy học có thể áp dụng cho bài toán dự đoán cấu trúc bậc cao của protein. Ở đây, khóa luận lựa chọn phương pháp máy học điểm hình đã cho kết quả khả quan với nhiều bài toán và có khả năng đạt kết quả tốt với bài toán dự đoán cấu trúc bậc cao của protein, đó là máy học SVM. Trong thực nghiệm thuộc phạm vi của khóa luận, bài toán dự đoán cấu trúc bậc cao của protein được xem là bài toán phân lớp, với các lớp chính là các nhãn của protein được xác định trước từ hai loại cây phân loại đó là SCOP và CATH.

Như đã giới thiệu ở phần trước, bài toán dự đoán cấu trúc bậc cao của protein dựa trên sự tương đồng của cấu trúc. Có nhiều cách để so sánh độ tương đồng của cấu trúc protein. Trong khóa luận này, sử dụng đồ thị để so sánh cấu trúc protein [35]. Mỗi đồ thị biểu diễn cho một protein, đặc trưng cho sự gần gũi trong chuỗi polypeptide. Khi đó, bài toán so sánh hai cấu trúc protein trở thành bài toán so sánh hai đồ thị.

Để xây dựng được mô hình dự đoán tốt, trước hết cần phải chọn đặc trưng tốt và phù hợp với bài toán. Dựa vào đồ thị của protein, khóa luận chọn vector phẳng làm vector đặc trưng cho từng protein. Trong toán học, vector phẳng là tập các giá trị riêng của ma trận được xây dựng từ đồ thị của protein.

Nội dung chương này chủ yếu trình bày lý thuyết để giải quyết bài toán dự đoán cấu trúc bậc cao của protein bằng cách xây dựng mô hình máy học SVM và mô hình K-Nearest Neighbors dựa trên các vector đặc trưng của protein là các vector phẳng. Cơ sở lý thuyết ở chương này sẽ là nền tảng cho phần thực nghiệm để đưa ra đánh giá về độ chính xác cũng như tính phù hợp của các phương pháp này với bài toán.

### 3.1 Biểu diễn cấu trúc 3D của protein bằng ma trận khoảng cách

Cấu trúc 3D của protein có thể được mô tả bằng ma trận khoảng cách hai chiều.

Ma trận khoảng cách  $DM_A$  của protein A với  $|N|$  residue là ma trận  $|N| \times |N|$ . Các thành phần  $DM_A[i, j]$  của ma trận khoảng cách được tính bằng khoảng cách từ atom  $C_\alpha$  thứ i cho đến atom  $C_\alpha$  thứ j ( $1 \leq i, j \leq |N|$ ), được ký hiệu là  $d_{ij}$ . Khoảng cách  $d_{ij}$  là khoảng cách Euclidean trong không gian 3D được định nghĩa như sau:

$$d_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2 + (z_i - z_j)^2} \quad (3.1)$$

Trong đó:

- $x_i, y_i, z_i$  là tọa độ không gian của atom  $C_\alpha$  thứ i.
- $x_j, y_j, z_j$  là tọa độ không gian của atom  $C_\alpha$  thứ j.

Nhận xét:

- Ma trận này là ma trận đối xứng nên  $d_{ij} = d_{ji}$ .
- Nếu  $i = j$  thì  $d_{ij} = 0$ .

|       | 1          | 2          | 3          | ... | $ N $        |
|-------|------------|------------|------------|-----|--------------|
| 1     | $d_{11}$   | $d_{12}$   | $d_{13}$   | ... | $d_{1 N }$   |
| 2     | $d_{21}$   | $d_{22}$   | $d_{23}$   | ... | $d_{2 N }$   |
| 3     | $d_{31}$   | $d_{32}$   | $d_{33}$   | ... | $d_{3 N }$   |
| :     | :          | :          | :          | :   | :            |
| $ N $ | $D_{ N 1}$ | $d_{ N 2}$ | $d_{ N 3}$ | ... | $d_{ N  N }$ |

Hình 3.1 – Biểu diễn ma trận khoảng cách 2D từ cấu trúc 3D của protein

### 3.2 Biểu diễn cấu trúc 3D của protein bằng ma trận kè

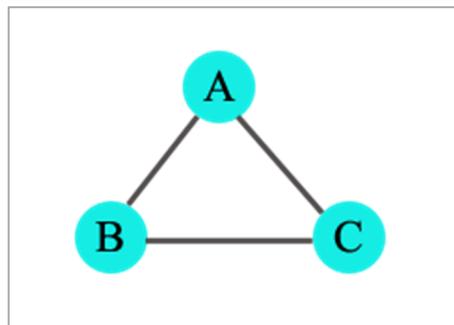
Xét đơn đồ thị vô hướng  $G = (V, E)$ , với tập đỉnh  $V = \{1, 2, \dots, n\}$ , tập cạnh  $E = \{e_1, e_2, \dots, e_m\}$ . Ta gọi ma trận kè (adjacency matrix) của đồ thị  $G$  là ma trận

$$A = \{a_{ij} : i, j = 1, 2, \dots, n\}$$

Với các phần tử được xác định theo qui tắc sau đây:

- $a_{ij} = 0$ , nếu  $(i, j) \notin E$
- $a_{ij} = 1$ , nếu  $(i, j) \in E, i, j = 1, 2, \dots, n$

Ví dụ: cho đồ thị vô hướng được biểu diễn như sau:



Hình 3.2 – Đồ thị vô hướng  $G$

Từ đồ thị vô hướng  $G$  ta biểu diễn ma trận kè như sau:

|   | A | B | C |
|---|---|---|---|
| A | 0 | 1 | 1 |
| B | 1 | 0 | 1 |
| C | 1 | 1 | 0 |

Ma trận kè là một ma trận  $n \times n$ , trong đó  $n$  là số đỉnh của đồ thị. Nếu có một cạnh nào đó nối đỉnh  $v_i$  với đỉnh  $v_j$  thì phần tử  $a_{ij} = 1$ , ngược lại có giá trị là 0.

Các tính chất của ma trận kè

- 1) Ma trận kè của đồ thị vô hướng là ma trận đối xứng, tức là  $a[i,j] = a[j,i]$ .
- 2) Tổng các phần tử trên dòng  $i$  (cột  $j$ ) của ma trận kè chính bằng bậc của đỉnh  $i$  (đỉnh  $j$ ).

Để xây dựng ma trận kè từ ma trận khoảng cách, theo các nhà khoa học [14], [34], nếu khoảng cách ( $d_{ij}$ ) giữa hai atom  $C_\alpha$  i và j thuộc ngưỡng giá trị  $6.5 \leq d_{ij} \leq 8.5$  (Angstrom) thì chúng được xem là kè nhau. Như vậy, sẽ có một cạnh nối giữa hai atom i và atom j.

Tóm lại, các phần tử ma trận kè của protein được xây dựng theo công thức sau:

$$a_{ij} = \begin{cases} 1, & \text{if } 6.5 \leq d_{ij} \leq 8.5 \\ 0, & \text{otherwise} \end{cases} \quad (3.2)$$

### 3.3 Phổ của đồ thị

Ma trận kè của đồ thị vô hướng có tính đối xứng, do đó có một tập đầy đủ các giá trị riêng (eigenvalue) và cơ sở vector riêng (eigenvector). Tập các giá trị riêng của đồ thị gọi là phổ của đồ thị (vector phổ). Dưới đây, khóa luận sẽ trình bày cách tìm vector phổ từ ma trận kè của protein.

#### Cách tìm vector phổ:

Cho A là ma trận vuông cấp n ( $A \in M_n(R)$ )

$$\begin{pmatrix} a_{11} & \dots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \dots & a_{nn} \end{pmatrix}$$

Khi đó

- Đa thức bậc n của biến  $\lambda$ :

$$P_A(\lambda) = \det(A - \lambda I) = \begin{vmatrix} a_{11} - \lambda & \dots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \dots & a_{nn} - \lambda \end{vmatrix}$$

$$= (-1)^n \lambda^n + a_{n-1} \lambda^{n-1} + \dots + a_1 \lambda_1 + a_0 \quad (3.3)$$

Gọi là đa thức đặc trưng của ma trận A

- Các nghiệm thực của đa thức đặc trưng  $P_A(\lambda)$  gọi là giá trị riêng của ma trận.

- Nếu  $\lambda_0$  là một giá trị riêng của A thì  $\det(A - \lambda_0 I) = 0$ . Do đó hệ phương trình thuần nhất

$$(A - \lambda_0 I) \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix} \quad (3.4)$$

có vô số nghiệm. Không gian nghiệm của hệ (3.4) gọi là không gian con riêng của ma trận A ứng với giá trị riêng  $\lambda_0$ . Các vector khác không là nghiệm của hệ (3.4) gọi là các vector riêng của ma trận A ứng với giá trị riêng  $\lambda_0$ .

Lấy ví dụ ở mục 3.2 để tìm vector phô.

$$A = \begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix}$$

$$\text{Ta có } P_A \lambda = \begin{vmatrix} -\lambda & 1 & 1 \\ 1 & -\lambda & 1 \\ 1 & 1 & -\lambda \end{vmatrix} = -\lambda^3 + 3\lambda + 2$$

Vậy đa thức đặc trưng của ma trận A là  $P_A \lambda = -\lambda^3 + 3\lambda + 2$

$$P_A \lambda = 0 \Leftrightarrow -\lambda^3 + 3\lambda + 2 = 0 \Leftrightarrow (\lambda + 1)^2(2 - \lambda) = 0 \Leftrightarrow \lambda = -1 \text{ (kép)}, \lambda = 2$$

Vậy ma trận A có ba giá trị riêng  $\lambda = -1, \lambda = -1, \lambda = 2$

Do đó, phô đồ thị của ma trận A là (2, -1, -1)

Lưu ý: phô đồ thị được sắp xếp theo chiều giảm dần của giá trị riêng.

### 3.4 Mô hình K-Nearest Neighbors (K-NN)

#### 3.4.1 Khái niệm chung

Phương pháp K-Nearest Neighbors (K-NN) [29] là phương pháp khá nổi tiếng theo hướng tiếp cận thống kê đã được nghiên cứu trong nhiều năm qua. K-NN được đánh giá là một trong những phương pháp phân loại có kết quả khá tốt.

Ý tưởng chính của phương pháp này là khi cần dự đoán lớp cho một đối tượng mới, thuật toán sẽ xác định khoảng cách (thường là khoảng cách Euclidean) giữa đối tượng cần phân lớp với tất cả các đối tượng trong tập dữ liệu huấn luyện để tìm ra k đối tượng gần nhất, gọi là K-Nearest Neighbor (k-láng giềng gần nhất), sau đó dùng các khoảng cách này để đánh trọng số cho tất cả các lớp cần phân loại vào. Khi đó, trọng số của một lớp chính là tổng nghịch đảo bình phương của tất cả các khoảng cách ở trên của các đối tượng trong k lảng giềng có cùng lớp. Lớp nào không xuất hiện trong k lảng giềng sẽ có trọng số bằng không. Sau đó các lớp sẽ được sắp xếp theo giá trị trọng số giảm dần và các lớp có trọng số cao sẽ được chọn làm lớp của đối tượng cần phân loại.

### 3.4.2 Cơ sở của phương pháp K-Nearest Neighbors

Gọi D là tập dữ liệu huấn luyện và một đối tượng cần phân loại là z, với một vector đặc trưng (là vector phẳng) và lớp của protein này chưa được biết. Thuật toán K-NN được mô tả như sau:

**Thuật toán k-NN**

**Input:** D là tập dữ liệu huấn luyện, đối tượng cần phân loại là z đặc trưng bởi một vector (vector phẳng) và L là tập các nhãn của đối tượng.

**Output:**  $c_z \in L$ , z là lớp dự đoán.

**foreach** đối tượng  $y \in D$  **do**

| Tính khoảng cách giữa hai đối tượng z và y  $d(z,y)$

**end**

Chọn  $N \subseteq D$ , là tập k đối tượng gần nhất với z;

$$c_z = \arg \max \sum_{y \in N} w_i \times I(v = \text{class}(c_y))$$

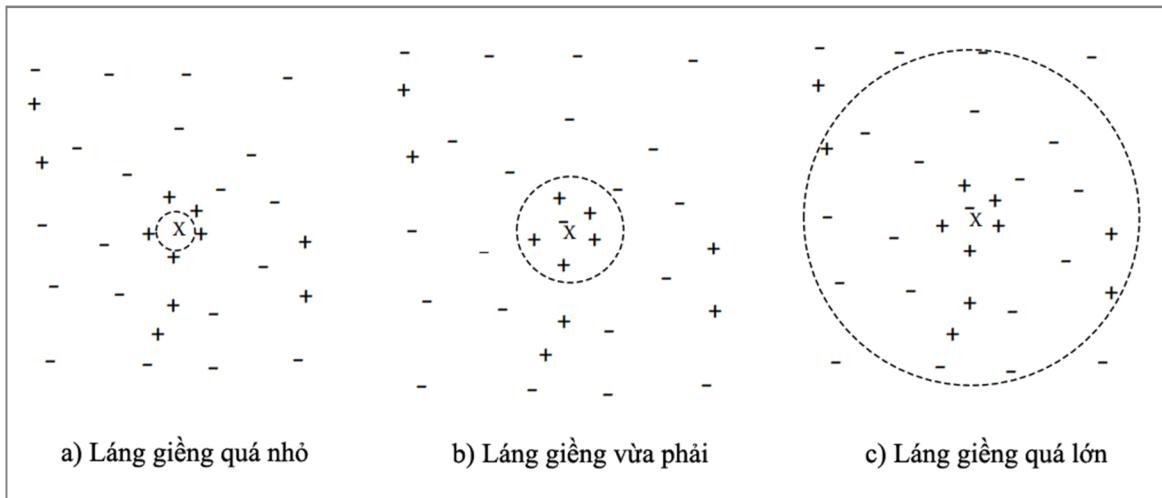
Với  $I(.)$  có giá trị là 1 nếu biểu thức bên trong đúng, và bằng 0 nếu ngược lại.

$$\text{Và } w = 1/d(z,y)^2$$

### 3.4.3 Chọn k

Chọn k bằng bao nhiêu là một vấn đề quan trọng của thuật toán K-NN. Trong trường hợp  $k = 1$ , ta có thuật toán lóng giềng gần nhất. Việc chọn giá trị k phải được tiến hành một cách cẩn thận. Nếu chọn giá trị k quá lớn thì có quá nhiều điểm với nhiều lớp khác nhau. Mặt khác, nếu ta chọn giá trị k quá nhỏ thì sẽ không đủ thông tin để dự đoán. Trường hợp dự đoán này sẽ không có tính khả thi.

Trong nhiều mô hình, người ta đã đưa ra một số giá trị k tỏ ra khá tốt với nhiều bài toán khác nhau, một trong số đó là cách chọn  $k = \sqrt{n}$ . Trong bài toán này, sau nhiều lần kiểm thử khác nhau, khóa luận chọn  $k = 20$ .



Hình 3.3 – K-Nearest Neighbor với các giá trị của k là quá nhỏ, vừa và quá lớn

### 3.4.4 Nhận xét

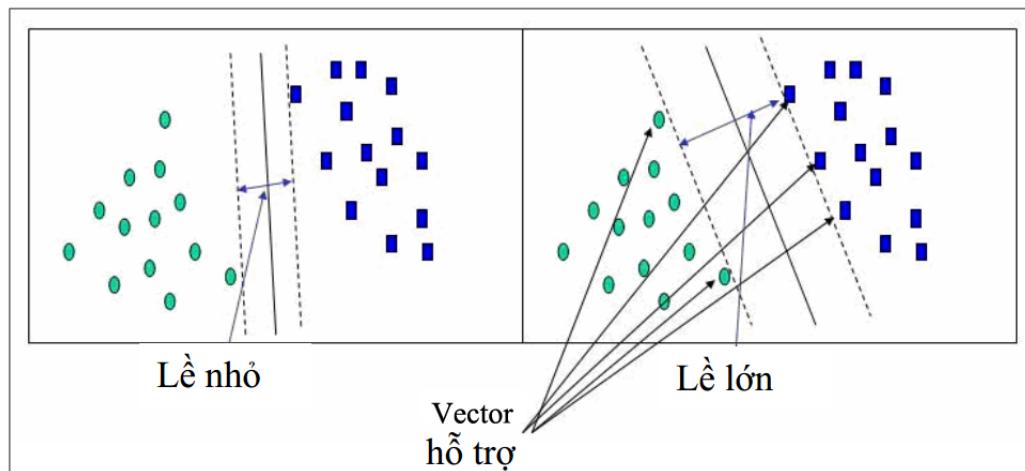
- **Ưu điểm**
  - Dễ dàng trong việc cài đặt thuật toán bởi tính dễ hiểu.
  - Có thể áp dụng cho nhiều loại khoảng cách độ đo như: Euclidean, Manhattan, Cosine...
- **Nhược điểm**
  - Khó khăn trong việc lựa chọn giá trị k.
  - Trường hợp có dữ liệu nhiều thì việc phân loại protein là không tốt.

### 3.5 Mô hình Support Vector Machine (SVM)

#### 3.5.1 Khái niệm chung

Phương pháp SVM (Support Vector Machine) [26], [27], [28] ra đời từ lý thuyết học thống kê do Vapnik và Chervonekis xây dựng vào năm 1992, và có nhiều tiềm năng phát triển về mặt lý thuyết cũng như ứng dụng trong thực tế. SVM là tập các phương pháp học có giám sát (supervised learning) được sử dụng để phân lớp dữ liệu. Phương pháp SVM có tính tổng quát cao nên có thể áp dụng cho nhiều loại bài toán nhận dạng và phân loại. Đây là phương pháp phân loại protein rất hiệu quả.

Ý tưởng chính của phương pháp là cho trước một tập huấn luyện biểu diễn trong không gian vector (trong bài toán này là các vector phẳng của protein). Như vậy, rõ ràng sẽ có nhiều cách có thể chia không gian này thành hai nữa riêng biệt, hình 3.4 cho ta một trường hợp về cách chia.

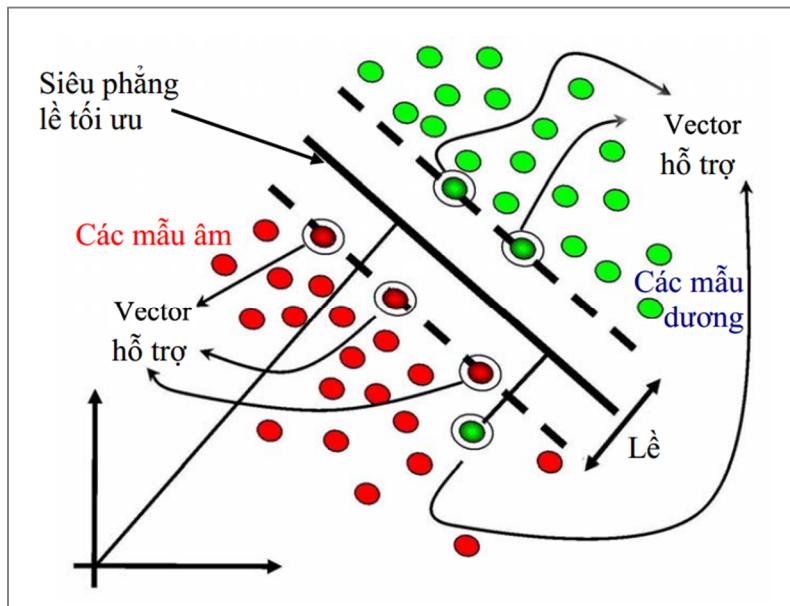


Hình 3.4 – Hai cách chia không gian vector thành hai nữa riêng biệt

Phương pháp SVM sẽ tìm ra một siêu mặt phẳng  $h$  (siêu phẳng) quyết định tốt nhất để chia các điểm trong không gian thành hai lớp riêng biệt tương ứng, tạm gọi là lớp âm (-) và lớp dương (+). Chất lượng của siêu phẳng này được quyết định bởi khoảng cách (được gọi là lè) của điểm dữ liệu gần nhất của mỗi lớp đến mặt phẳng này. Khoảng cách lè càng lớn thì xác suất của việc phân lớp sai sẽ càng nhỏ, tức là

càng có sự phân chia tốt các điểm ra thành hai lớp, như vậy, ta sẽ đạt được kết quả phân lớp tốt. Theo [28], bộ phân lớp SVM là mặt siêu phẳng phân tách các mẫu dương ra khỏi các mẫu âm với độ chênh lệch cực đại, trong đó độ chênh lệch – còn gọi là lề – xác định bằng khoảng cách giữa các mẫu dương và các mẫu âm gần mặt siêu phẳng nhất. Mặt siêu phẳng này được gọi là siêu phẳng lề tối ưu.

Tóm lại, mục đích của SVM là tìm được khoảng cách lề lớn nhất và lỗi tách sai bé nhất để tạo kết quả phân lớp tốt. Hình 3.5 cho ta mô tả trực quan về phương pháp SVM.

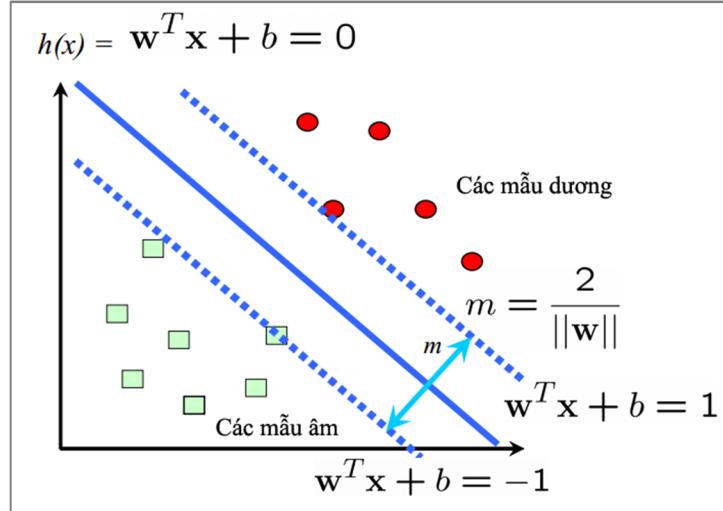


Hình 3.5 – Mặt siêu phẳng tách các mẫu dương ra khỏi các mẫu âm

### 3.5.2 Cơ sở của phương pháp SVM

Cho tập mẫu  $\{X_i, y_i\}$ ,  $i = 1, \dots, N$ ,  $y_i \in \{-1, +1\}$ ,  $X_i \in R^n$ , với  $y_i$  là một số nguyên xác định lớp  $X_i$ ,  $N$  là số lượng tập mẫu. Trong đó mẫu là các vector đối tượng được phân loại vào các mẫu dương và mẫu âm. Các mẫu dương là các mẫu  $X_i$  thuộc lĩnh vực quan tâm và được gián nhãn  $y_i = +1$ . Ngược lại, các mẫu âm là các mẫu  $X_i$  không thuộc lĩnh vực quan tâm và được gián nhãn  $y_i = -1$ .

Mục tiêu của phương pháp phân lớp SVM là tìm một siêu phẳng phân cách sao cho khoảng cách lè (margin) giữa hai lớp đạt cực đại (hình 3.6).



Hình 3.6 – Siêu phẳng tách với khoảng cách lè cực đại

Từ bài toán đặt ra là tìm mặt phẳng phân tách  $h(x) = w^T x + b$ . Trong đó  $w$  là vector pháp tuyến của siêu phẳng,  $b$  đóng vai trò là độ dịch. Bộ phân loại SVM được định nghĩa như sau:

$$f(x) = \text{sign}(w^T x + b) = \begin{cases} 1, & w^T x + b \geq 0 \\ -1, & w^T x + b < 0 \end{cases} \quad (3.5)$$

Nếu  $f(x) = +1$  thì  $x$  thuộc về lớp dương (lĩnh vực đang được quan tâm).

Nếu  $f(x) = -1$  thì  $x$  thuộc về lớp âm (các lĩnh vực khác).

Máy học SVM là một họ các mặt siêu phẳng phụ thuộc vào các tham số  $w$  và  $b$ . Mục tiêu của phương pháp SVM là ước lượng  $w$  và  $b$  để cực đại hóa lè giữa các lớp dữ liệu dương và âm.

Trường hợp nếu tập dữ liệu huấn luyện là khả tách tuyến tính, ta có các ràng buộc sau:

- $w^T x_i + b \geq +1$  nếu  $y_i = +1$  (3.6)

- $w^T x_i + b \leq -1$  nếu  $y_i = -1$  (3.7)

Hai mặt phẳng có phương trình là  $w^T x + b = \pm 1$  được gọi là các mặt siêu phẳng hỗ trợ (các đường nét đứt trên hình 3.6).

Để xây dựng một mặt siêu phẳng lè tối ưu, ta phải giải bài toán quy hoạch toàn phương sau:

Cực đại hóa:

$$\sum_{i=1}^N \alpha_i - \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j y_i y_j x_i^T x_j \quad (3.8)$$

Với các ràng buộc:

$$\bullet \quad \alpha_i \geq 0 \quad (3.9)$$

$$\bullet \quad \sum_{i=1}^N \alpha_i y_i = 0 \quad (3.10)$$

Trong đó, các hệ số Lagrange  $\alpha_i$ ,  $i = 1, 2, \dots, N$ , là các biến cần được tối ưu hóa.

Vector  $w$  sẽ được tính từ các nghiệm của bài toán toàn phương nô trên như sau:

$$w = \sum_{i=1}^N \alpha_i y_i x_i \quad (3.11)$$

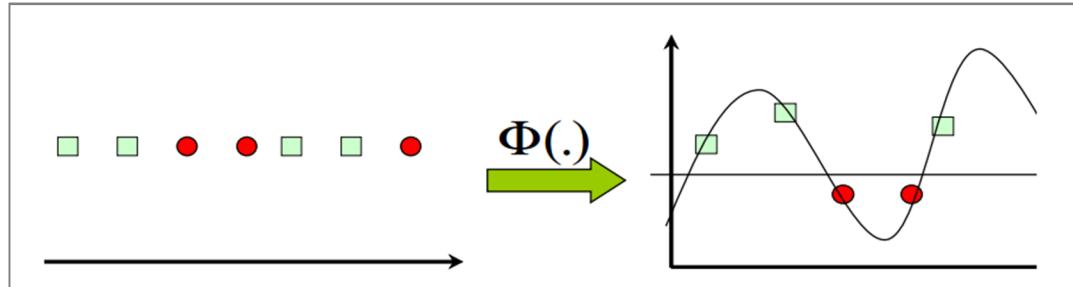
Để xác định độ dịch  $b$ , ta chọn mẫu  $x_i$  sao cho với  $\alpha_i > 0$ , sau đó sử dụng điều kiện Karush-Kuhn-Tucker (KKT) như sau:

$$\alpha_i [y_i (w^T x_i + b) - 1] = 0 \quad (3.12)$$

Các mẫu  $x_i$  tương ứng với  $\alpha_i > 0$  là những mẫu nằm gần mặt siêu phẳng quyết định nhất và được gọi là các vector hỗ trợ.

Trường hợp nếu tập dữ liệu là không khả tách tuyến tính (khả tách phi tuyến), thì ta có thể giải quyết theo cách như sau: sử dụng một ánh xạ phi tuyến  $\Phi$  để ánh xạ các điểm dữ liệu đầu vào sang một không gian mới (không gian đặc trưng) có số chiều cao hơn. Trong không gian này, các điểm dữ liệu trở thành khả tách tuyến tính, hoặc có thể phân tách với ít lỗi hơn so với trường hợp sử dụng không gian ban

đầu. Một mặt quyết định tuyến tính trong không gian mới sẽ tương ứng với một mặt quyết định phi tuyến trong không gian ban đầu.



Hình 3.7 – Chuyển không gian ban đầu vào không gian đặc trưng  
Khi đó, bài toán quy hoạch toàn phương ban đầu sẽ trở thành:

Cực đại hóa:

$$\sum_{i=1}^N \alpha_i - \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j y_i y_j k(x_i x_j) \quad (3.13)$$

Với các ràng buộc:

- $0 \leq \alpha_i \leq C$  (3.14)

- $\sum_{i=1}^N \alpha_i y_i = 0$  (3.15)

Trong đó  $k$  là một hàm nhân thỏa mãn:

$$k(x_i x_j) = \Phi(x_i)^T \cdot \Phi(x_j) \quad (3.16)$$

Một trong những ưu điểm của phương pháp SVM so với các thuật toán học khác là mặt siêu phẳng quyết định chỉ phụ thuộc và các vector hỗ trợ, khi các điểm khác bị xóa đi thì thuật toán vẫn cho kết quả giống ban đầu. Như vậy, tất cả các dữ liệu huấn luyện đều được dùng để tối ưu hóa thuật toán.

Một vấn đề lớn được đặt ra là phương pháp SVM có thể chia dữ liệu làm hai lớp, tuy nhiên đối với bài toán dự đoán cấu trúc bậc cao của protein thì số lớp tương ứng loại protein mà ta cần xác định luôn lớn hơn hai. Câu hỏi đặt ra là liệu phương pháp SVM có phù hợp để giải quyết bài toán đặt ra hay không?. Để giải quyết vấn

đề này, thường thì dữ liệu lớn hơn hai lớp sẽ được xử lý bằng phương pháp pairwise, tức là với dữ liệu chứa N lớp, ta sẽ xây dựng tất cả các cặp của hai lớp khác nhau, tổng số sẽ là  $N(N-1)/2$  cặp. Lớp được chọn sẽ được xác định dựa trên cơ sở đánh giá kết quả của  $N(N-1)/2$  lần phân lớp.

### 3.5.3 Vai trò của hàm kernel

Hàm kernel đóng một vai trò quan trọng trong bài toán phân loại sử dụng phương pháp SVM, nó tạo ra ma trận kernel tóm tắt tất cả dữ liệu. Việc chọn một hàm kernel dùng trong mô hình là rất khó bởi vì nó còn phụ thuộc vào dữ liệu xây dựng mô hình.

Dưới đây là một số hàm kernel phổ biến được khóa luận chọn:

- Linear:  $K(x, y) = (x^T y)^d$
- Polynomial:  $K(x, y) = (x^T y + 1)^d$
- Gaussia Radial Basis Function:  $K(x, y) = \exp\left(-\frac{\|x - y\|^2}{2\delta^2}\right)$
- Sigmoid:  $K(x, y) = \tanh(kx^T y + \theta)$

### 3.5.4 Huấn luyện SVM

Huấn luyện SVM thực chất là việc giải bài toán quy hoạch toàn phương SVM [24], [27]. Các phương pháp số giải bài toán quy hoạch này yêu cầu phải lưu trữ một ma trận có kích thước bằng bình phương của số lượng mẫu huấn luyện. Trong những bài toán thực tế thì điều này là không khả thi vì thông thường kích thước của tập dữ liệu huấn luyện thường rất lớn (có thể lên tới hàng chục nghìn mẫu). Nhiều thuật toán khác nhau được phát triển để giải quyết vấn đề nêu trên. Những thuật toán này dựa trên việc phân rã tập dữ liệu huấn luyện thành những nhóm dữ liệu. Điều đó có nghĩa là bài toán quy hoạch toàn phương lớn được phân rã thành các bài

toán quy hoạch toàn phương với kích thước nhỏ hơn. Sau đó, những thuật toán này kiểm tra các điều kiện KKT (Karush Kuhn Tucker) để xác định phương án tối ưu.

Khóa luận tập trung vào nghiên cứu thuật toán huấn luyện SVM tối ưu hóa tuần tự cực tiểu (Sequential Minimal Optimization – SMO) [24]. Thuật toán này sử dụng tập dữ liệu huấn luyện (còn gọi là tập làm việc) có kích thước nhỏ nhất bao gồm hai hệ số Lagrange. Bài toán quy hoạch toàn phương nhỏ nhất phải gồm hai hệ số Lagrange vì các hệ số Lagrange phải thỏa mãn ràng buộc đẳng thức 3.15. Phương pháp SMO cũng có một số heuristic cho việc chọn hai hệ số Lagrange để tối ưu hóa ở mỗi bước. Mặc dù có nhiều bài toán quy hoạch toàn phương con hơn so với các phương pháp khác, mỗi bài toán con này được giải rất nhanh dẫn đến bài toán được giải một cách nhanh chóng.

### 3.5.5 Nhận xét

- **Ưu điểm:**
  - Là phương pháp phân lớp nhanh và có hiệu quả cao.
  - SVM giải quyết vấn đề dữ liệu nhiễu rất tốt.
  - Rất có hiệu quả để giải quyết bài toán dữ liệu có số chiều lớn. Phù hợp với đặc trưng của protein mà khóa luận đã chọn.
- **Nhược điểm:**
  - Để đạt được kết quả phân loại tốt cần chọn hàm kernel phù hợp.
  - Yêu cầu phải lập đi lập lại quá trình huấn luyện đối với bài toán nhiều lớp.

## 3.6 Tổng kết chương 3

Chương này đã xem xét cách giải quyết bài toán dự đoán cấu trúc bậc cao của protein bằng hai cách là xây dựng mô hình SVM và K-NN. Đặc biệt là cách sử dụng phương pháp đồ thị để so sánh độ tương đồng của từng cấu trúc protein. Dựa vào đồ thị để trích rút các vector phô (vector đặc trưng) của từng protein. Với mô

hình phân lớp bằng phương pháp SVM, trình bày cách giải bài toán quy hoạch toàn phương dựa vào thuật toán SMO.

Dựa vào mô hình SVM, đối với bài toán dự đoán cấu trúc protein, thì hàm kernel Gaussia được chọn đầu tiên bởi vì có nhiều lý do. Hàm kernel Gaussia phi tuyến ánh xạ tập mẫu học vào không gian có kích thước rộng lớn. Hơn nữa hàm kernel tuyến tính là trường hợp đặc biệt của hàm kernel phi tuyến. Cuối cùng, hàm kernel Gaussia có độ phức tạp ít hơn.

Chương tiếp theo sẽ trình bày các thử nghiệm thực tế các nghiên cứu lý thuyết trên tập dữ liệu cụ thể được lấy từ PDB và đánh giá kết quả đạt được.

## CHƯƠNG 4

# XÂY DỰNG HỆ THỐNG CHƯƠNG TRÌNH

Như đã biết, protein là những thành phần chức năng của cơ thể, sự hình thành và phát triển chức năng mới dựa trên cơ sở thành phần của cấu trúc, hay nói cách khác giữa chức năng và thành phần cấu trúc của protein có liên quan mật thiết với nhau.

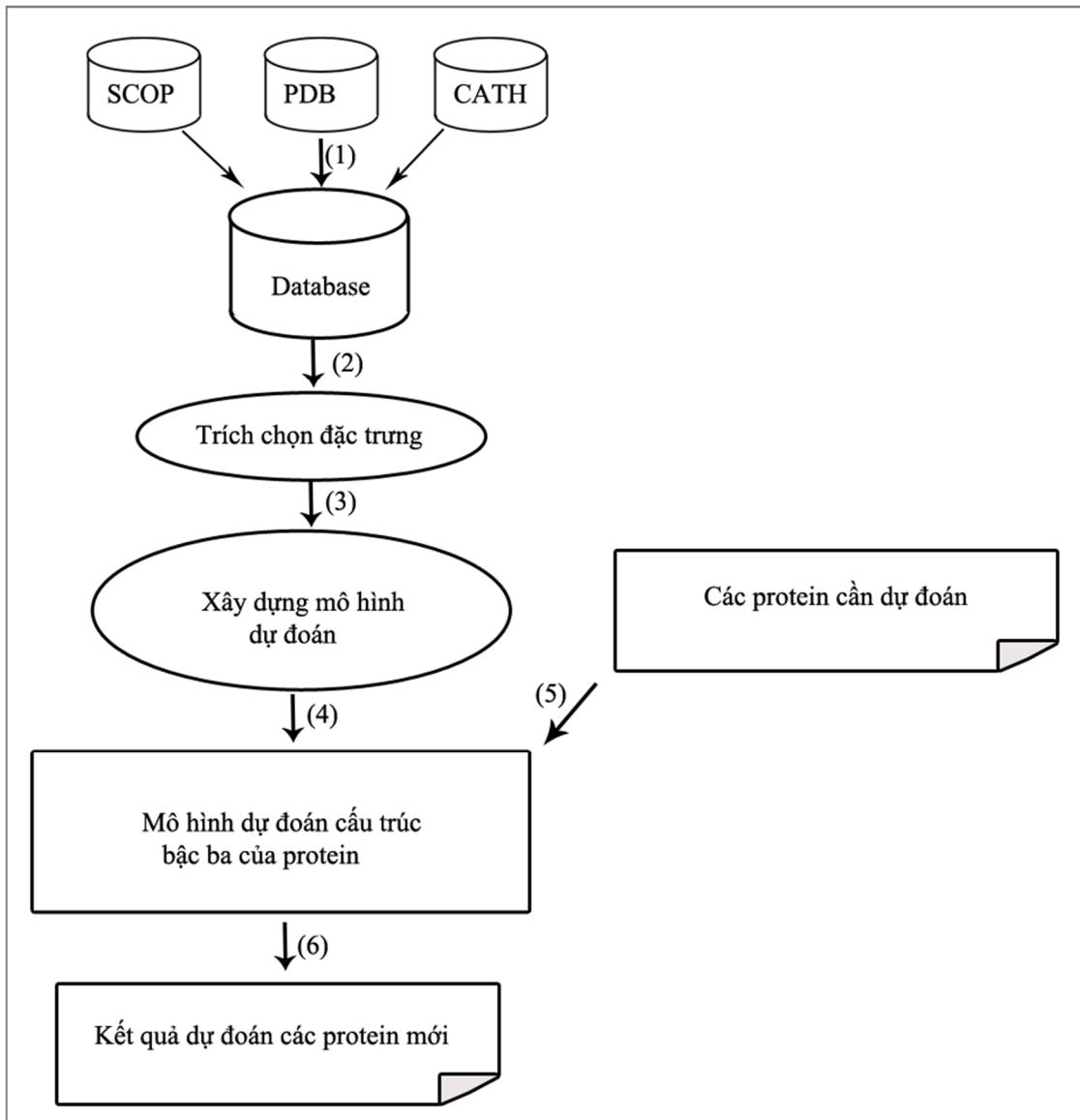
Dựa trên cơ sở lý thuyết ở chương 3, khóa luận tiến hành thực nghiệm áp dụng mô hình phân lớp SVM và K-NN cho bài toán dự đoán cấu trúc bậc cao của protein dựa trên hai cây phân loại là SCOP và CATH. Từ kết quả thu được, khóa luận đưa ra một số kết quả so sánh đã đạt được cũng như một số nhận xét sơ bộ về ưu và nhược điểm của phương pháp này.

#### **4.1 Hệ thống chương trình**

Hệ thống chương trình được chia thành bốn nhiệm vụ chính:

- 1) Xây dựng mô hình phân tích các đặc trưng của protein là các vector phô, tiến hành lưu trữ thông tin các cây phân loại như SCOP và CATH cũng như đối tượng protein.
- 2) Xây dựng mô hình phân loại đối với hai mô hình là SVM và K-NN. Mô hình học dữ liệu từ tập dữ liệu huấn luyện.
- 3) Kiểm thử độ chính xác của mô hình bằng tập dữ liệu kiểm thử, kết quả trả về là độ chính xác và thời gian kiểm thử của mô hình.
- 4) Dự đoán cấu trúc protein mới, kết quả trả về là các lớp và phần trăm tương ứng mà protein đó có thể thuộc vào.

#### 4.1.1 Quy trình thực hiện tổng quát



Hình 4.1 – Quy trình thực hiện tổng quát

Hệ thống chương trình thực hiện qua các giai đoạn sau:

- Thu gom dữ liệu từ các cơ sở dữ liệu có liên quan thông qua mạng Internet. Các cơ sở dữ liệu liên quan bao gồm cơ sở dữ liệu PDB dùng để lấy thông tin về đối tượng protein, cơ sở dữ liệu cây phân loại SCOP và CATH dùng để phân loại protein. Từ các cơ sở dữ liệu này, phân tích tìm

- hiểu một số đặc trưng của sinh học cần thiết của đối tượng protein, lập cơ sở dữ liệu và rút trích các thông tin có liên quan vào cơ sở dữ liệu cục bộ.
2. Xây dựng các đặc trưng của đối tượng protein là các vector phô. Khi đó, mỗi protein sẽ được đại diện bởi vector phô tương ứng hỗ trợ cho việc xây dựng mô hình dự đoán.
  3. Từ các đặc trưng trích chọn, đặc biệt là các vector phô, hệ thống tiến hành xây dựng mô hình dự đoán SVM và K-NN. Thuật toán SMO được sử dụng để giải quyết bài toán SVM.
  4. Sau khi xây dựng được mô hình dự đoán, hệ thống sẽ có cơ chế lưu lại các thông tin, các tham số của mô hình máy học để phục vụ cho những lần kiểm thử sau mà không cần phải chạy lại chương trình.
  5. Khi đã có được mô hình dự đoán cấu trúc của protein. Ta tiến hành kiểm thử độ tin cậy (độ chính xác) của mô hình. Bằng cách chọn ngẫu nhiên các protein để kiểm thử, những protein này được chọn sao cho khác với những protein đã được huấn luyện.
  6. Là giai đoạn xử lý cho ra kết quả của mô hình dự đoán cấu trúc protein.

#### 4.1.2 Mô hình dự đoán cho cấu trúc protein mới

Thuật toán dự đoán cấu trúc protein mới:

Đầu vào: protein cần dự đoán P.

Đầu ra: các lớp và tỉ lệ phần trăm tương ứng mà protein có thể thuộc vào.

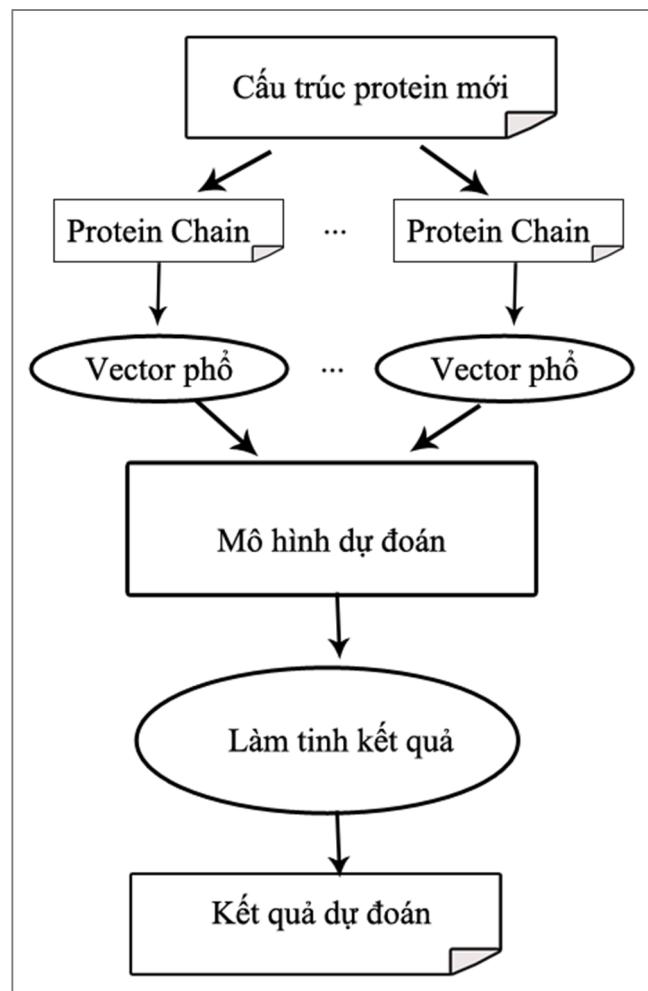
Bước 1: Đọc các thông tin chính của protein P. Chủ yếu là các thông tin về atom của protein.

Bước 2: Tách các atom của protein theo từng chain. Khi đó, ta xem các atom sau khi tách như là một protein con  $P_i$ , hệ thống sẽ dự đoán theo những protein  $P_i$  này.

Bước 3: Tiếp theo, hệ thống sẽ tìm vector phô cho từng protein  $P_i$ . Sau khi có được vector phô, ta phải chuyển vector phô này về cùng số chiều với các vector phô khi huấn luyện tập dữ liệu học.

Bước 4: Sử dụng mô hình SVM hay K-NN để phân lớp dựa vào đặc trưng là vector phô.

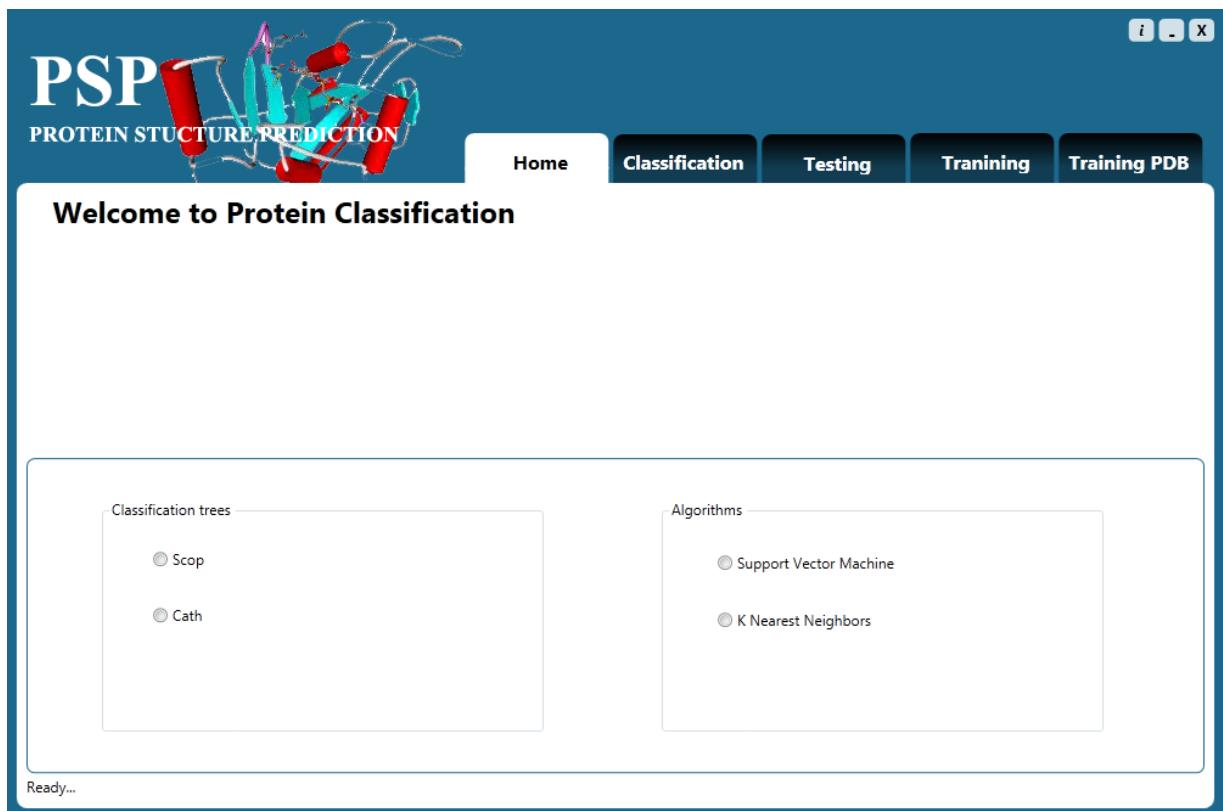
Bước 5: Làm tinh kết quả dự đoán.



Hình 4.2 - Mô hình dự đoán cho cấu trúc protein mới

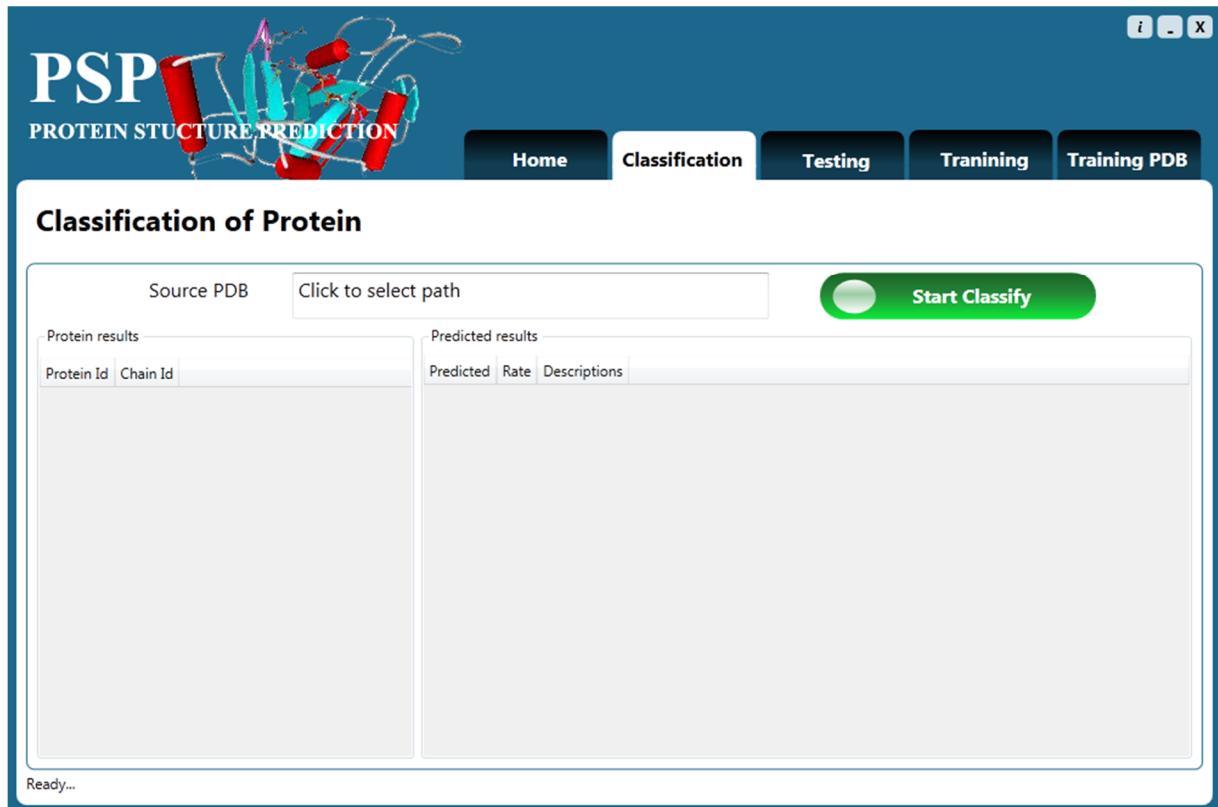
#### 4.1.3 Một số màn hình chức năng tiêu biểu

##### 4.1.3.1 Màn hình khởi động chương trình



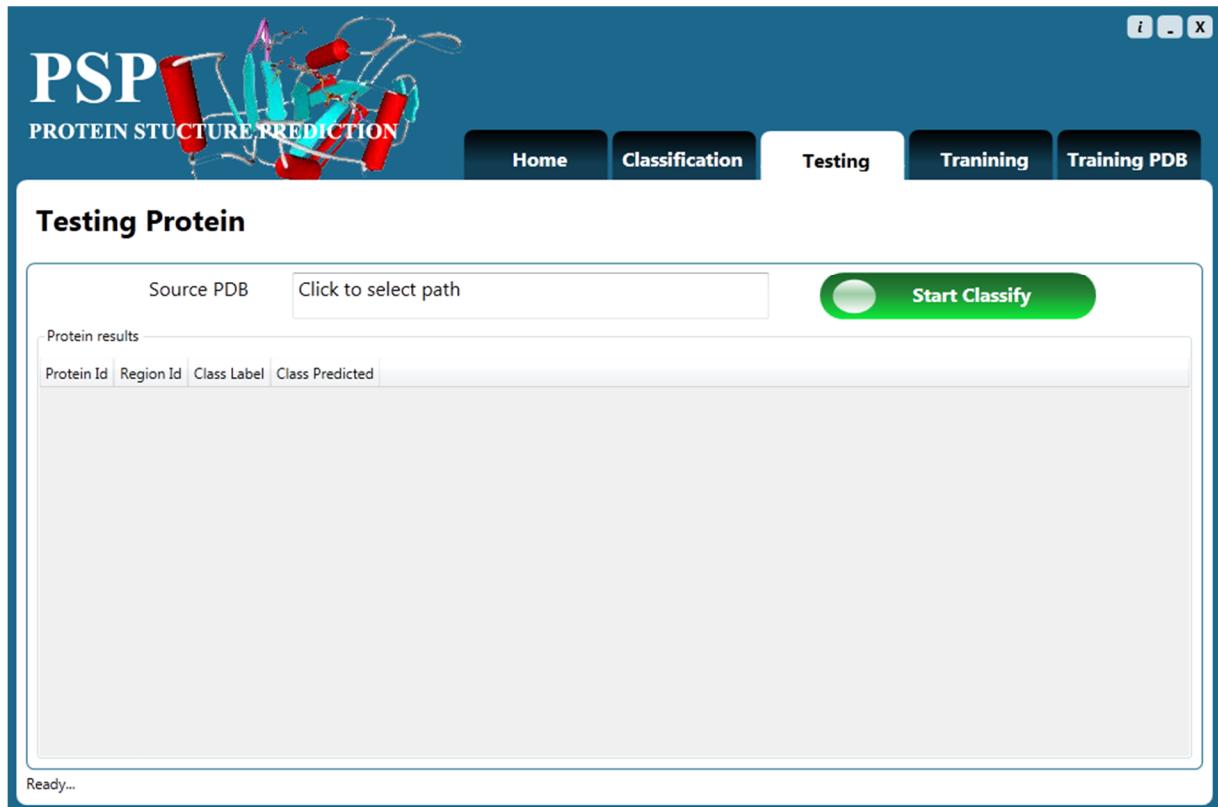
Hình 4.3 – Màn hình khởi động chương trình

#### 4.1.3.2 Màn hình dự đoán cấu trúc protein mới



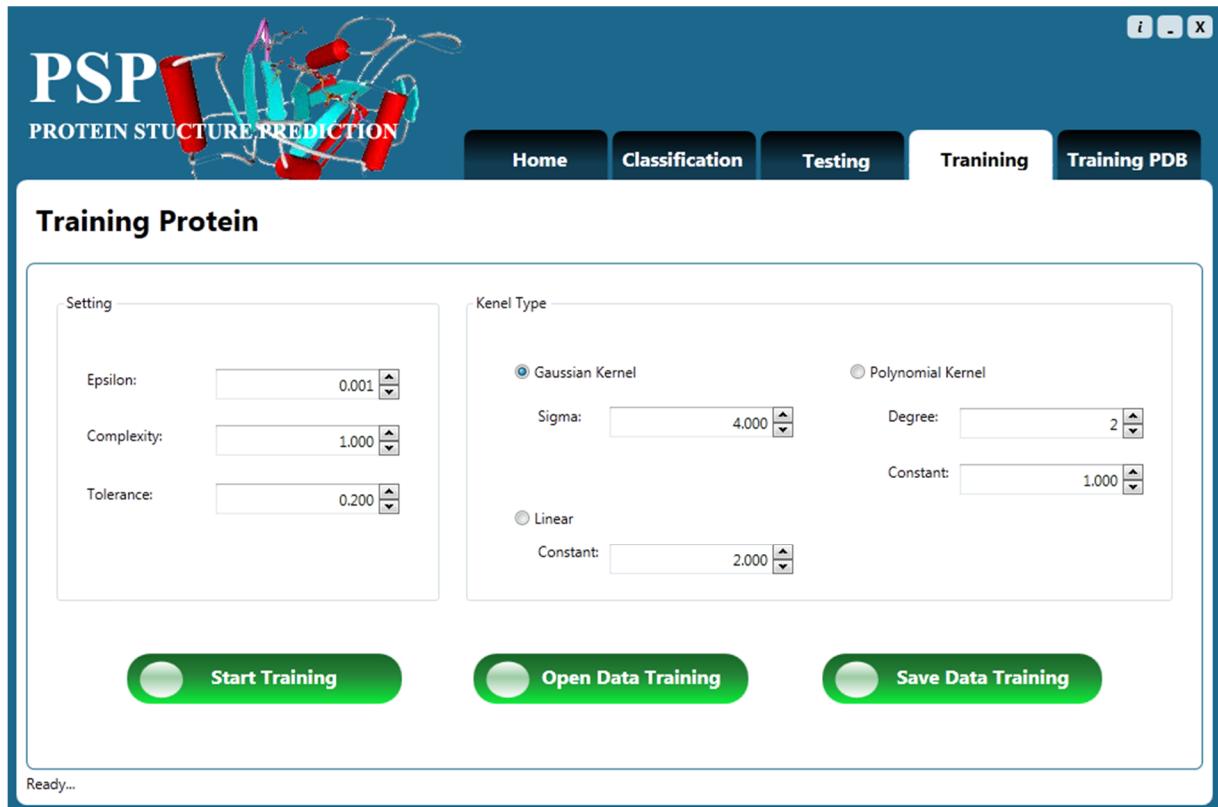
Hình 4.4 – Màn hình dự đoán cấu trúc protein mới

#### 4.1.3.3 Màn hình kiểm thử



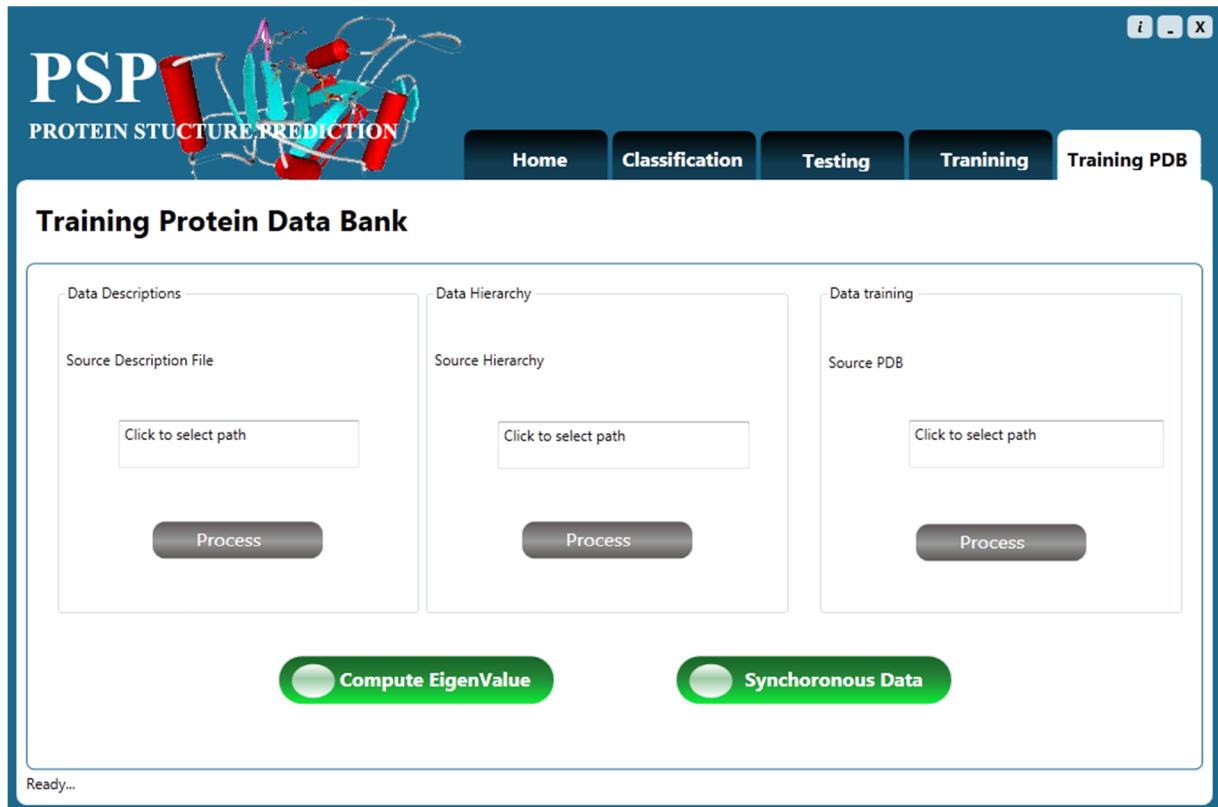
Hình 4.5 – Màn hình kiểm thử

#### 4.1.3.4 Màn hình máy học SVM



Hình 4.6 – Màn hình máy học SVM

#### 4.1.3.5 Màn hình học dữ liệu từ PDB, SCOP và CATH



Hình 4.7 – Màn hình học dữ liệu từ PDB, SCOP và CATH

## 4.2 Kết quả thực hiện

### 4.2.1 Các tham số đánh giá phân lớp

- Độ chính xác của kết quả: Đây là một trong những yếu tố quan trọng nhất cần phải xem xét để đánh giá độ tốt của một mô hình. Đối với các mô hình, độ chính xác của dữ liệu đầu ra được tính bằng công thức:

$$P = \frac{\text{correct}}{\text{correct} + \text{incorrect}}$$

- Tốc độ xử lý: Trong một số tình huống, tốc độ phân lớp được xem như là một yếu tố quan trọng. Một bộ phận lớp với độ chính xác 92% có thể được chuộng hơn bộ phận lớp có độ chính xác 95% nhưng chậm hơn 100 lần để có thể đưa ra được kết quả.
- Dễ hiểu: Một bộ phận lớp dễ hiểu sẽ tạo cho người sử dụng tin tưởng hơn vào hệ thống, đồng thời cũng giúp cho người sử dụng tránh được việc hiểu lầm kết quả của một luật được đưa ra bởi hệ thống.
- Thời gian để học: Vấn đề này đặc biệt nghiêm trọng khi hệ thống được sử dụng trong các môi trường thay đổi thường xuyên, điều đó yêu cầu hệ thống phải học rất nhanh một luật phân lớp hoặc nhanh chóng điều chỉnh một luật đã học cho phù hợp với thực tế.

### 4.2.2 Môi trường cài đặt

Hệ thống được triển khai dựa trên ngôn ngữ C# và được thực hiện chủ yếu trên nền hệ điều hành Windows 7, SQL Server 2008, bộ công cụ Visual Studio 2010. Các kết quả được thực hiện trên máy tính xách tay có bộ xử lý Core i3, 2.13GHz và 2GB Ram.

#### 4.2.3 Mô tả tập dữ liệu huấn luyện và tập dữ liệu kiểm thử

Khóa luận đã thực hiện thu thập dữ liệu protein từ ngân hàng PDB ở địa chỉ <ftp://ftp.wwpdb.org/pub/pdb/data/structures/divided/pdb/>. Kho lưu trữ PDB thành lập nhằm mục đích phục vụ cộng đồng trên thế giới, các nhà nghiên cứu, các nhà giáo dục, các sinh viên nghiên cứu sinh học. Do đó, có thể lấy về hoàn toàn miễn phí. Có hai loại định dạng phổ biến đó là ‘.ent’ hay ‘.pdb’. Với tổng cộng protein được lấy về là hơn 20,400 protein khác nhau, sau khi đã loại trừ với điều kiện là các protein thuộc cùng một nhóm phải lớn hơn 100 protein để có đủ đặc trưng và phải thuộc cây phân loại SCOP và CATH thì mới được chọn. Số lượng được chọn là 4294 protein đối với cây phân loại SCOP, còn đối với CATH là 5649 protein. Sở dĩ số lượng protein được chọn cho CATH nhiều hơn SCOP là do cây phân loại CATH là phân loại theo bán tự động nên số lượng protein đã được biết đến nhiều hơn cây phân loại SCOP là phân loại bằng tay. Sau khi lấy dữ liệu về được chia thành hai tập: một tập dùng để huấn luyện với tỉ lệ là 80% và tập còn lại dùng để kiểm thử được lấy trong tổng số dữ liệu được lấy về. Bảng 4.1 là bảng thống kê chi tiết dữ liệu được sử dụng trong hệ thống.

Bảng 4.1 – Bảng thống kê dữ liệu huấn luyện và dữ liệu kiểm thử

| Cây phân loại    | SCOP       |          | CATH       |          |
|------------------|------------|----------|------------|----------|
|                  | Huấn luyện | Kiểm thử | Huấn luyện | Kiểm thử |
| Số lượng protein | 3504       | 790      | 4819       | 830      |
| Tổng protein     | 4294       |          | 5649       |          |

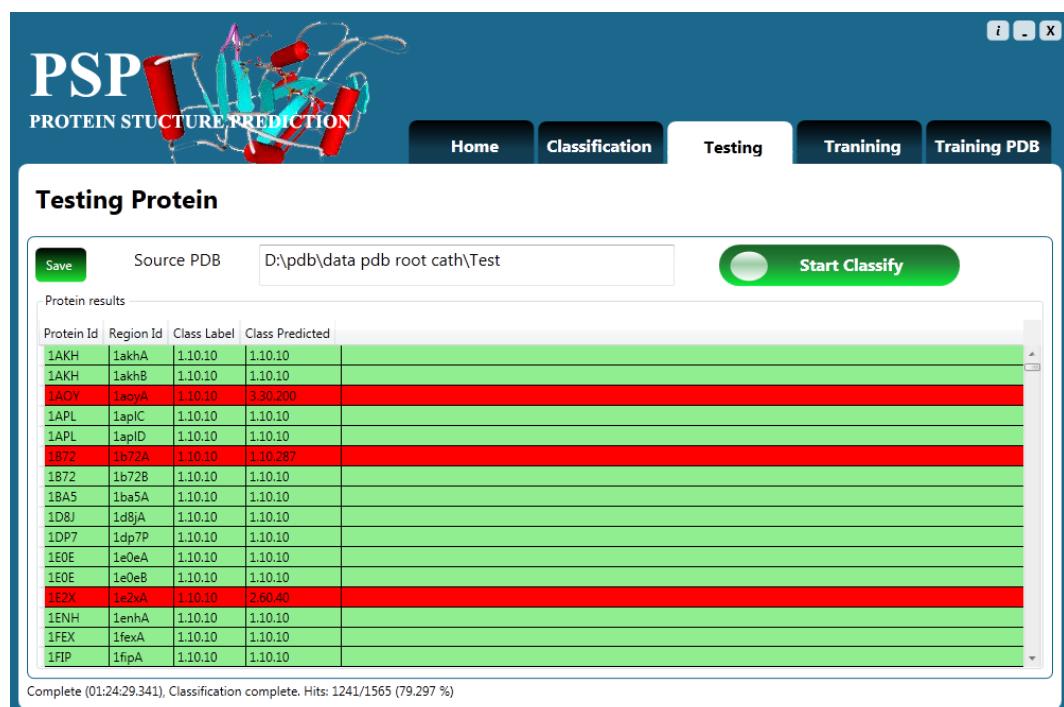
Tập dữ liệu huấn luyện là tập dữ liệu được dùng cho việc xây dựng mô hình dự đoán, tập dữ liệu kiểm thử là tập dữ liệu dùng để kiểm tra độ chính xác của mô hình.

#### 4.2.4 Kết quả thực hiện kiểm thử

Trong quá trình kiểm thử trên nhiều tập, kết quả thực nghiệm cho thấy rằng một mô hình có mạnh hay không còn phụ thuộc vào nguồn dữ liệu huấn luyện. Dữ liệu huấn luyện được thu thập càng phong phú, khả năng dự đoán của mô hình với dữ liệu mới càng tăng.

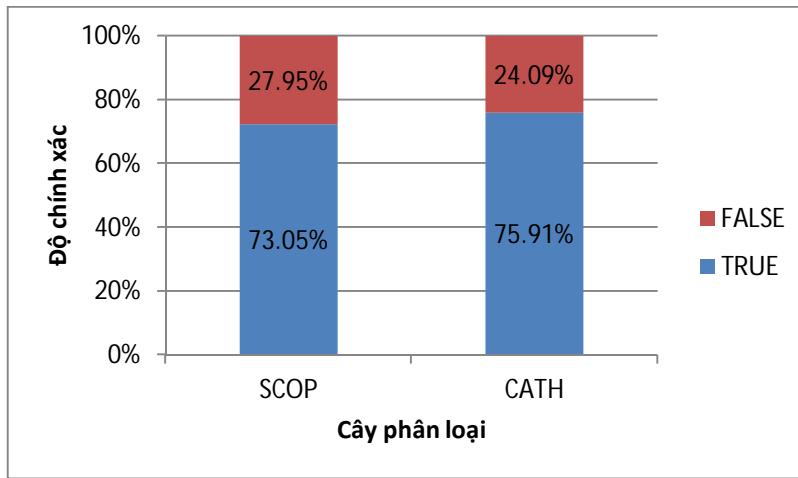
- a) Dữ liệu đầu vào: Là tập dữ liệu kiểm thử.
- b) Dữ liệu đầu ra: Hệ thống đưa ra lớp mà chương trình dự đoán với lớp mà protein thuộc vào tương ứng với từng chain của protein. Kết quả được hiển thị trực tiếp trên màn hình dự đoán và có thể lưu vào tập tin.
- c) Kết quả

Sau khi tiến hành chạy chương trình với bộ dữ liệu kiểm thử trên hai cây phân loại SCOP và CATH với cả hai mô hình SVM và K-NN thì hệ thống cho được kết quả khá tốt. Kết quả này rất có ý nghĩa, nó thể hiện sức mạnh thực sự của mô hình với tập đặc trưng là vector phô.



Hình 4.8 – Kết quả kiểm thử cấu trúc protein theo cây phân loại CATH.

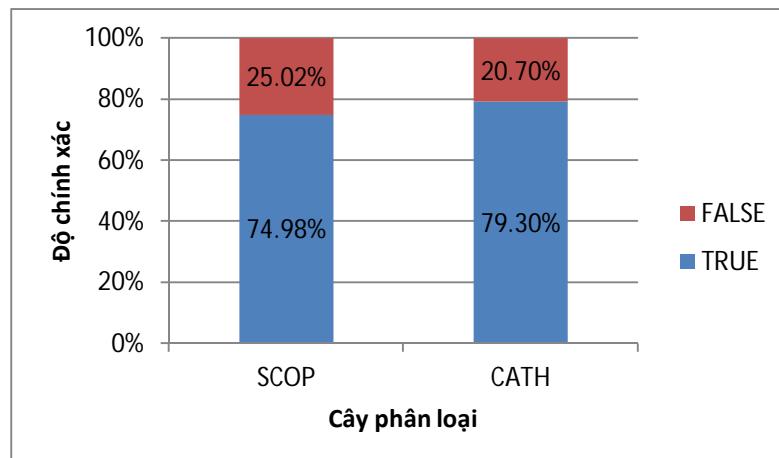
- Đổi với mô hình K-NN



Hình 4.9 – Biểu đồ so sánh kết quả kiểm thử trên SCOP và CATH đổi với mô hình K-NN

Biểu đồ trên biểu diễn sự so sánh kết quả kiểm thử đổi với hai cây phân loại SCOP và CATH dựa vào mô hình là K-NN. Đổi với cây phân loại CATH thì hệ thống cho kết quả khá tốt tỉ lệ dự đoán đúng là 75.91%, còn đổi với SCOP là 73.05%. Thời gian kiểm thử đổi với cây phân loại SCOP là 85 phút, còn đổi với CATH là khoảng 120 phút.

- Đổi với mô hình SVM



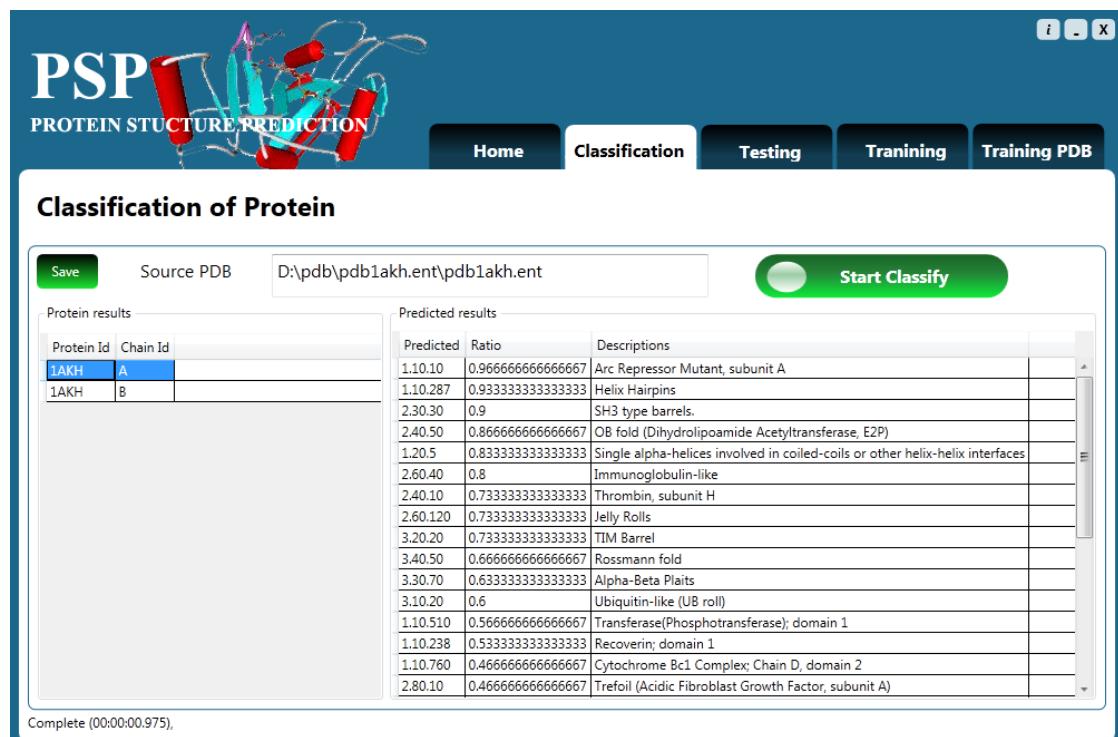
Hình 4.10 – Biểu đồ so sánh kết quả kiểm thử trên SCOP và CATH đổi với mô hình SVM

Biểu đồ trên biểu diễn sự so sánh kết quả kiểm thử đối với hai cây phân loại SCOP và CATH dựa vào mô hình là SVM. Đối với cây phân loại CATH thì hệ thống cho kết quả khá tốt tỉ lệ dự đoán đúng là 79.30%. Thời gian kiểm thử đối với SCOP là 72 phút, và với CATH là 85 phút.

#### 4.2.5 Kết quả dự đoán cấu trúc protein mới

- Dữ liệu đầu vào: Các thông tin chính về protein cần dự đoán cấu trúc. Thông tin về protein chủ yếu được lấy từ PDB.
- Dữ liệu đầu ra: Các lớp mà hệ thống dự đoán cấu trúc protein thuộc vào tương ứng với tỉ lệ phần trăm mà protein có khả năng thuộc vào lớp đó. Kết quả được hiển thị trực tiếp trên màn hình dự đoán và có thể lưu vào tập tin.
- Kết quả

Chức năng này dùng để dự đoán cấu trúc của những protein chưa biết. Nhằm mục đích hỗ trợ cho các nhà sinh học trong việc tìm hiểu chức năng của protein.



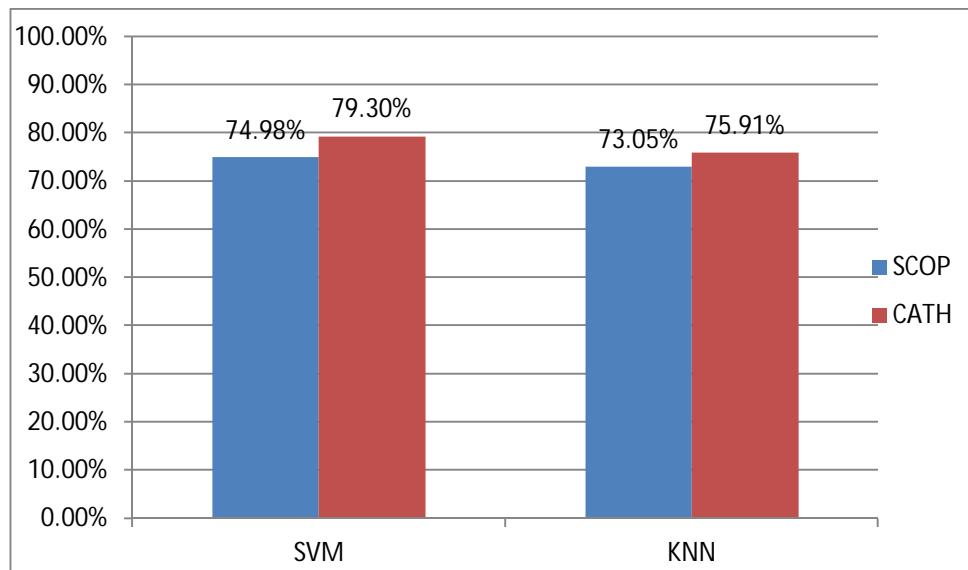
Hình 4.11 – Kết quả dự đoán cấu trúc protein mới theo cây phân loại CATH

### 4.3 Tổng kết chương 4

Trong chương này, khóa luận đã trình bày các chức năng của hệ thống, những kết quả của quá trình huấn luyện cũng như kiểm thử chức năng dự đoán cấu trúc bậc cao của protein dựa vào mô hình phân lớp là SVM và K-NN. Tập dữ liệu huấn luyện của mô hình được xây dựng một cách khá công phu với 3504 protein đối với cây phân loại SCOP còn đối với CATH là 4819 protein. Dựa vào những kết quả thực nghiệm đã đạt được, nó càng khẳng định tính đúng đắn ở ba chương trước.

Để tận dụng cấu hình máy hiện tại, khóa luận sử dụng một thư viện AForge [39], là thư viện mã nguồn mở giúp tận dụng khả năng xử lý đồng thời của bộ xử lý. Với bộ xử lý Core i3 có bốn nhân, sẽ chạy đồng thời bốn tiểu trình, với mỗi tiểu trình sẽ sử dụng một nhân. Giúp cho việc chạy thuật toán được nhanh hơn. Thời gian để huấn luyện thuật toán SVM trên bộ dữ liệu SCOP là 1 giờ 35 phút, còn trên bộ dữ liệu CATH là 4 giờ 29 phút.

Hình 4.12 là biểu đồ so sánh các kết quả đã đạt được qua hai mô hình phân lớp là SVM và K-NN cho bài toán dự đoán cấu trúc bậc cao của protein.



Hình 4.12 – Biểu đồ so sánh kết quả giữa hai mô hình SVM và K-NN

Qua kết quả thực nghiệm và những nhận xét đánh giá, với khả năng dự đoán của mô hình SVM lên tới 79.30% đối với CATH và 74.98% đối với SCOP với thời gian chấp nhận được, bài toán dự đoán cấu trúc bậc cao của protein rất có triển vọng để áp dụng vào các ứng dụng trong tương lai.

## **CHƯƠNG 5**

### **KẾT LUẬN VÀ KIẾN NGHỊ**

## 5.1 Kết luận

Trong khuôn khổ của một khóa luận tốt nghiệp đại học, nội dung nghiên cứu tập trung tìm hiểu bài toán dự đoán cấu trúc bậc cao của protein và cách giải quyết. Tuy chưa đạt được một kết quả đặc biệt vượt trội, nhưng hy vọng khóa luận sẽ góp phần đem lại lợi ích cho cộng đồng nghiên cứu về vấn đề dự đoán cấu trúc bậc cao của protein.

Khóa luận đã trình bày một số lý thuyết về protein, chức năng cũng như tầm quan trọng của việc dự đoán cấu trúc protein. Đồng thời, khóa luận cũng đã tìm hiểu và áp dụng hai cây phân loại phổ biến hiện nay như SCOP và CATH.

Tìm hiểu mô hình so sánh cấu trúc của protein dựa trên mô hình tương đồng. Để so sánh độ tương đồng giữa các cấu trúc protein, khóa luận sử dụng phương pháp đồ thị.

Dựa trên lý thuyết đã tìm hiểu được, khóa luận đề xuất một phương pháp máy học là SVM dựa trên đặc trưng là các vector phẳng. Tiến hành cài đặt và thực nghiệm áp dụng trên hai mô hình phân lớp là SVM và K-NN cho cả hai cây phân loại SCOP và CATH trên bộ dữ liệu mà khóa luận đã chọn và cùng cách lấy đặc trưng của mỗi protein để đưa ra so sánh một cách khách quan.

Từ thực nghiệm cho thấy kết quả khả quan của hướng tiếp cận dựa trên hai mô hình là SVM và K-NN. Với đặc trưng của protein là các vector phẳng cho thấy cả hai phương pháp đều cho kết quả rất đáng chú ý. Trong đó, phương pháp SVM luôn cho độ chính xác cao nhất trong tất cả các thực nghiệm.

Với kết quả thực nghiệm cho thấy phương pháp máy học SVM đã áp dụng cho kết quả khá tốt. Bộ phân loại protein dựa trên cây phân loại SCOP cho kết quả với độ chính xác 74.98% và dựa trên cây phân loại CATH cho kết quả với độ chính xác tốt nhất là 79.30% với cùng hàm kernel Gaussia. SVM có ưu thế về mặt thời gian huấn luyện khá tốt, tốc độ phân lớp cũng chấp nhận được. Kết quả thu được là khá

tương đồng với các nghiên cứu trên thế giới, điều này chứng tỏ tập đặc trưng mà khóa luận lựa chọn là phù hợp với bài toán.

Kết quả mà khóa luận đã đạt được, tuy chưa thật sự xuất sắc, nhưng cũng đã đạt được yêu cầu ban đầu đặt ra và đặt nền tảng cho những nghiên cứu tiếp theo.

## 5.2 Kiến nghị

Do còn nhiều hạn chế về thời gian và kiến thức, khóa luận còn một số vấn đề cần tiếp tục hoàn thiện và phát triển trong thời gian tới:

- Tiếp tục nghiên cứu kỹ hơn về lý thuyết các mô hình máy học, thay đổi các tham số, thuật toán hay hàm nhân được sử dụng khi áp dụng mô hình với hy vọng cải thiện kết quả tốt hơn.
- Tìm hiểu thêm các đặc trưng mới của protein, để có thể áp dụng nhiều mô hình máy học hơn và nhằm tăng độ chính xác của kết quả.
- Tìm hiểu và cài đặt vài mô hình dự đoán cấu trúc nữa để có kết quả so sánh với mô hình hiện tại.
- Tiếp tục phát triển và hoàn thiện chương trình với huy vọng có thể ứng dụng mô hình vào thực tế hiện tại của Việt Nam.

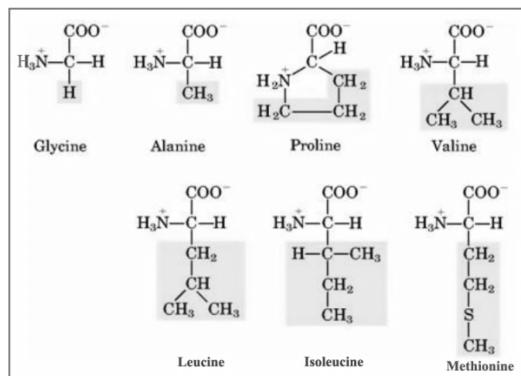
## PHỤ LỤC

### A. Phân loại amino acid

Hiện nay có nhiều cách khác nhau để phân loại amino acid, mỗi cách sắp xếp đều có ý nghĩa và mục đích riêng. Tuy nhiên, các cách này đều dựa trên cấu tạo hóa học hoặc một số tính chất của gốc R. Ví dụ: có cách chia các amino acid thành hai nhóm chính đó là nhóm mạch thẳng và nhóm mạch vòng. Trong nhóm mạch thẳng lại tùy theo sự có mặt của số nhóm carboxyl hay số nhóm amino mà chia ra thành các nhóm nhỏ, nhóm amino acid trung tính (chứa một nhóm COOH và một nhóm NH<sub>2</sub>); nhóm amino acid có tính kiềm (chứa một nhóm COOH và hai nhóm NH<sub>2</sub>). Trong nhóm mạch vòng lại chia ra thành nhóm đồng vòng hay dị vòng v.v... Cách khác lại dựa vào tính phân cực của gốc R chia các amino acid thành 4 nhóm: nhóm không phân cực hoặc kỵ nước, nhóm phân cực nhưng không tích điện, nhóm tích điện dương và nhóm tích điện âm.

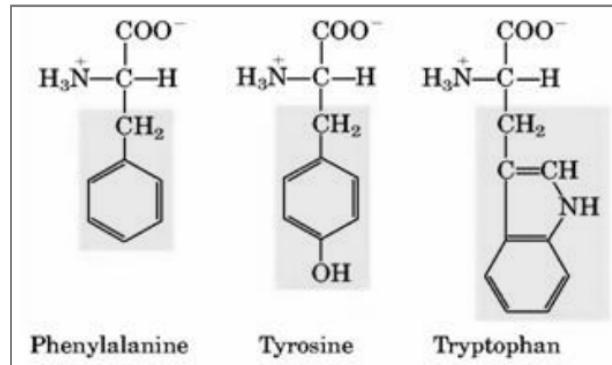
Ở đây, khóa luận này xin được giới thiệu cách phân loại các amino acid một cách chung nhất của PGS. TS. Cao Đăng Nguyên [1]. Theo cách này dựa vào gốc R các amino acid được chia làm năm nhóm:

Nhóm I: gồm bảy amino acid có gốc R không phân cực và kỵ nước, đó là: glycine, alanine, proline, valine, leucine, isoleucine và methionine.



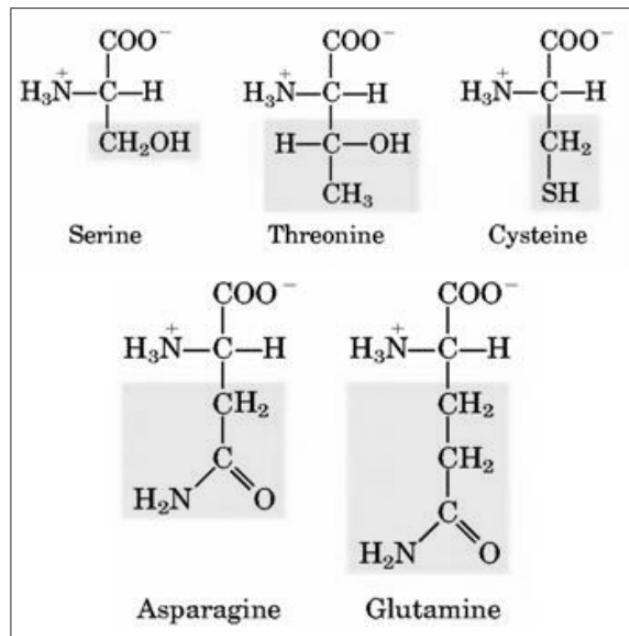
Hình A.1 – Công thức cấu tạo các amino acid nhóm I

Nhóm II: gồm ba amino acid có gốc R chứa nhân thơm, đó là phenylalanine, tyrosine và tryptophan.



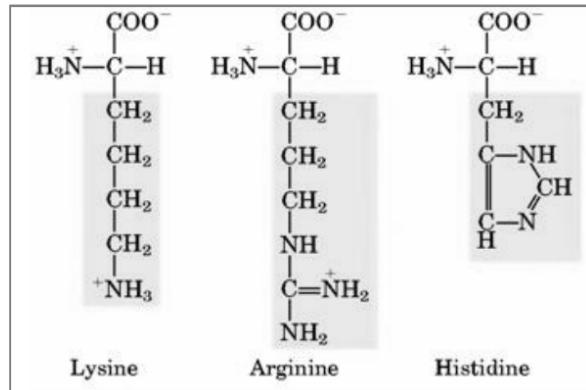
Hình A.2 – Công thức cấu tạo các amino acid nhóm II.

Nhóm III: gồm năm amino acid có gốc R phân cực, không tích điện, đó là serine, threonine, cysteine, asparagine và glutamine.



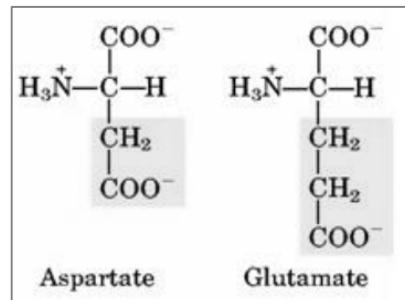
Hình A.3 – Công thức cấu tạo các amino acid nhóm III.

Nhóm IV: gồm ba amino acid có gốc R tích điện dương, đó là lysine, histidine và arginine.



Hình A.4 – Công thức cấu tạo các amino acid nhóm IV.

Nhóm V: gồm hai amino acid có gốc R tích điện âm, đó là aspartate và glutamate.



Hình A.5 – Công thức cấu tạo các amino acid nhóm V.

Hai mươi loại amino acid khác nhau được tìm thấy trong các protein có thể được viết tắt bằng hệ thống ba chữ cái (bảng A.1). Chữ viết tắt các amino acid thường là những chữ cái đầu của tên, nhưng một số amino acid đôi khi bắt đầu với cùng một chữ của bảng chữ cái. Chúng đặc biệt được sử dụng khi viết ra các trình tự protein. Các amide, asparagine và glutamine, tương đối không ổn định và bị phá vỡ một cách dễ dàng thành các axit tương ứng, là aspartate và glutamate.

Bảng A.1 – Hai mươi amino acid trong protein

| Tên amino acid | Tên viết tắt | Ký hiệu |
|----------------|--------------|---------|
| Alanine        | Ala          | A       |
| Arginine       | Arg          | R       |
| Asparagine     | Asn          | N       |
| Aspartic acid  | Asp          | D       |
| Cysteine       | Cys          | C       |
| Glutamic acid  | Glu          | E       |
| Glutamine      | Gln          | Q       |
| Glycine        | Gly          | G       |
| Histidine      | His          | H       |
| Isoleucine     | Ile          | I       |
| Leucine        | Leu          | L       |
| Lysine         | Lys          | K       |
| Methionine     | Met          | M       |
| Phenylalanine  | Phe          | F       |
| Proline        | Pro          | P       |
| Serine         | Ser          | S       |
| Threonine      | Thr          | T       |
| Tryptophan     | Trp          | W       |
| Tyrosine       | Tyr          | Y       |
| Valine         | Val          | V       |

## B. Một số kết quả dự đoán xuất ra dạng file.

### B.1 Kết quả kiểm thử protein

Sau khi tiến hành kiểm thử tập dữ liệu protein thì ta có thể xem trên giao diện chương trình hoặc xuất ra dưới dạng file để xem. Hình B.1 là một giao diện được xuất ra sau khi tiến hành kiểm thử bằng mô hình SVM trên cây phân loại là CATH.

| #      | Result test protein classification  |        |  |             |  |  |  |  |  |  |                     |
|--------|-------------------------------------|--------|--|-------------|--|--|--|--|--|--|---------------------|
| #      | Release date : 9/1/2012 10:05:48 PM |        |  |             |  |  |  |  |  |  |                     |
| #      | Algorithm type : KNearestNeighbors  |        |  |             |  |  |  |  |  |  |                     |
| #      | Classification tree type : Scop     |        |  |             |  |  |  |  |  |  |                     |
| #      | Proteinid                           | Region |  | Label       |  |  |  |  |  |  | Predicted           |
| #===== |                                     |        |  |             |  |  |  |  |  |  |                     |
| 1A0Z   | ----- 1a0zA -----                   |        |  | 46456.46457 |  |  |  |  |  |  | 46456.46457 => true |
| 1A0Z   | ----- 1a0zB -----                   |        |  | 46456.46457 |  |  |  |  |  |  | 46456.46457 => true |
| 1A0Z   | ----- 1a0zC -----                   |        |  | 46456.46457 |  |  |  |  |  |  | 46456.46457 => true |
| 1A0Z   | ----- 1a0zD -----                   |        |  | 46456.46457 |  |  |  |  |  |  | 46456.46457 => true |
| 1A3N   | ----- 1a3nA -----                   |        |  | 46456.46457 |  |  |  |  |  |  | 46456.46457 => true |
| 1A3N   | ----- 1a3nB -----                   |        |  | 46456.46457 |  |  |  |  |  |  | 46456.46457 => true |
| 1A3N   | ----- 1a3nC -----                   |        |  | 46456.46457 |  |  |  |  |  |  | 46456.46457 => true |
| .      | .                                   |        |  | .           |  |  |  |  |  |  | .                   |
| .      | .                                   |        |  | .           |  |  |  |  |  |  | .                   |
| 2BTS   | ----- 2btsA -----                   |        |  | 53931.56111 |  |  |  |  |  |  | 53931.56111 => true |
| 2C1B   | ----- 2c1bA -----                   |        |  | 53931.56111 |  |  |  |  |  |  | 53931.56111 => true |
| 2C6I   | ----- 2c6iA -----                   |        |  | 53931.56111 |  |  |  |  |  |  | 53931.56111 => true |
| 2C6M   | ----- 2c6mA -----                   |        |  | 53931.56111 |  |  |  |  |  |  | 53931.56111 => true |
| 2E9P   | ----- 2e9pA -----                   |        |  | 53931.56111 |  |  |  |  |  |  | 53931.56111 => true |
| 2ERK   | ----- 2erkA -----                   |        |  | 53931.56111 |  |  |  |  |  |  | 53931.56111 => true |
| 2ETM   | ----- 2etmA -----                   |        |  | 53931.56111 |  |  |  |  |  |  | 53931.56111 => true |
| 2ETM   | ----- 2etmB -----                   |        |  | 53931.56111 |  |  |  |  |  |  | 53931.56111 => true |
| 3BV3   | ----- 3bv3A -----                   |        |  | 53931.56111 |  |  |  |  |  |  | 53931.56111 => true |
| 3BYS   | ----- 3bysA -----                   |        |  | 53931.56111 |  |  |  |  |  |  | 53931.56111 => true |
| 3E87   | ----- 3e87A -----                   |        |  | 53931.56111 |  |  |  |  |  |  | 51349.51350         |
| 3E87   | ----- 3e87B -----                   |        |  | 53931.56111 |  |  |  |  |  |  | 51349.51350         |
| 3E92   | ----- 3e92A -----                   |        |  | 53931.56111 |  |  |  |  |  |  | 53931.56111 => true |
| #===== |                                     |        |  |             |  |  |  |  |  |  |                     |
| #      | Total of protein chain: 1347        |        |  |             |  |  |  |  |  |  |                     |
| #      | Number of prediction correct: 984   |        |  |             |  |  |  |  |  |  |                     |
| #      | Number of prediction incorrect: 363 |        |  |             |  |  |  |  |  |  |                     |
| #      | Accuracy: 73.051                    |        |  |             |  |  |  |  |  |  |                     |

Hình B.1 – Kết quả kiểm thử xuất ra dưới dạng file

Hình B.1 mô tả kết quả kiểm thử của mô hình SVM sử dụng cây phân loại CATH. Dưới đây là bảng mô tả chi tiết về kết quả kiểm thử.

- (1): Release date là ngày kiểm thử (định dạng theo tháng/ngày/năm).
- (2): Algorithm type là loại mô hình đang được sử dụng (SVM hoặc K-NN).
- (3): Classification tree type là cây phân loại đang được sử dụng để dự đoán (SCOP hoặc CATH).
- (4): Mô tả mã protein (idcode) tương ứng trong SCOP hay CATH.
- (5): Vùng dự đoán protein, như là một chain trong protein (ký hiệu: idcode\_chain).
- (6): Lớp thực mà trong SCOP hay CATH phân loại, theo định dạng là Class.Fold nếu cây phân loại SCOP được sử dụng hoặc Classs.Architecture.Topology nếu cây phân loại CATH được sử dụng.
- (7): Lớp mà chương trình dự đoán, cũng theo định dạng như (6).
- (8): Nếu lớp thực tê và lớp mà chương trình dự đoán là như nhau của cùng một protein chain thì kết quả dự đoán là chính xác tức là true, ngược lại là false.
- (9): Tổng số protein chain kiểm thử.
- (10): Tổng số protein chain dự đoán đúng.
- (11): Tổng số protein chain dự đoán sai.
- (12): Tỉ lệ phần trăm dự đoán đúng.

## B.2 Kết quả dự đoán protein mới

Chức năng dự đoán protein nhằm hỗ trợ cho các nhà sinh học. Chức năng này dùng để dự đoán khi cần phân tích chức năng của protein mà chưa biết chức năng của chúng. Nhiệm vụ của hệ thống là đưa ra các lớp dự đoán với tỉ lệ phần trăm mà lớp đó có thể thuộc vào. Hình B.2 là giao diện được xuất ra dưới dạng file.

|  |    |  |  |  |  |  |  |  |  |
|--|----|--|--|--|--|--|--|--|--|
| # Result of protein classification                                       |    |  |  |  |  |  |  |  |  |
| # Release date : 8/29/2012 11:32:54 PM                                   |    |  |  |  |  |  |  |  |  |
| # Algorithm type : SVM   |    |  |  |  |  |  |  |  |  |
| # Classification tree type : Cath  |    |  |  |  |  |  |  |  |  |
| =====  |    |  |  |  |  |  |  |  |  |
| Protein Id : 1AKH  | 1  |  |  |  |  |  |  |  |  |
| Protein chain : A  | 2  |  |  |  |  |  |  |  |  |
| Predicted :  | 3  |  |  |  |  |  |  |  |  |
| ProteinId  | 4  |  |  |  |  |  |  |  |  |
| Chainid  | 5  |  |  |  |  |  |  |  |  |
| Predicted  | 6  |  |  |  |  |  |  |  |  |
|  | 7  |  |  |  |  |  |  |  |  |
|  | 8  |  |  |  |  |  |  |  |  |
|  | 9  |  |  |  |  |  |  |  |  |
|  | 10 |  |  |  |  |  |  |  |  |
|  |    |  |  |  |  |  |  |  |  |
| 1AKH ----- A ----- 1.10.10 ----- 0.9666666666666667 ----- Arc Repressor  |    |  |  |  |  |  |  |  |  |
| 1AKH ----- A ----- 1.10.287 ----- 0.9333333333333333 ----- Helix Hairpin |    |  |  |  |  |  |  |  |  |
| 1AKH ----- A ----- 2.30.30 ----- 0.9 ----- SH3 type barrels.             |    |  |  |  |  |  |  |  |  |
| 1AKH ----- A ----- 2.40.50 ----- 0.8666666666666667 ----- OB fold (Dihyd |    |  |  |  |  |  |  |  |  |
| 1AKH ----- A ----- 1.20.5 ----- 0.8333333333333333 ----- Single alpha-he |    |  |  |  |  |  |  |  |  |
| 1AKH ----- A ----- 2.60.40 ----- 0.8 ----- Immunoglobulin-like           |    |  |  |  |  |  |  |  |  |
| 1AKH ----- A ----- 2.40.10 ----- 0.7333333333333333 ----- Thrombin, subu |    |  |  |  |  |  |  |  |  |
| 1AKH ----- A ----- 2.60.120 ----- 0.7333333333333333 ----- Jelly Rolls   |    |  |  |  |  |  |  |  |  |
| 1AKH ----- A ----- 3.20.20 ----- 0.7333333333333333 ----- TIM Barrel     |    |  |  |  |  |  |  |  |  |
| 1AKH ----- A ----- 3.40.50 ----- 0.6666666666666667 ----- Rossmann fold  |    |  |  |  |  |  |  |  |  |
| 1AKH ----- A ----- 3.30.70 ----- 0.6333333333333333 ----- Alpha-Beta Pla |    |  |  |  |  |  |  |  |  |
| 1AKH ----- A ----- 3.10.20 ----- 0.6 ----- Ubiquitin-like (UB roll)      |    |  |  |  |  |  |  |  |  |
| 1AKH ----- A ----- 1.10.510 ----- 0.5666666666666667 ----- Transferase(P |    |  |  |  |  |  |  |  |  |
| 1AKH ----- A ----- 1.10.238 ----- 0.5333333333333333 ----- Recoverin; do |    |  |  |  |  |  |  |  |  |
| 1AKH ----- A ----- 1.10.760 ----- 0.4666666666666667 ----- Cytochrome Bc |    |  |  |  |  |  |  |  |  |
| 1AKH ----- A ----- 2.80.10 ----- 0.4666666666666667 ----- Trefoil (Acidi |    |  |  |  |  |  |  |  |  |
| 1AKH ----- A ----- 1.10.150 ----- 0.4333333333333333 ----- DNA polymeras |    |  |  |  |  |  |  |  |  |
| 1AKH ----- A ----- 2.40.128 ----- 0.4 ----- Lipocalin                    |    |  |  |  |  |  |  |  |  |
| 1AKH ----- A ----- 2.40.70 ----- 0.3333333333333333 ----- Cathepsin D, s |    |  |  |  |  |  |  |  |  |

Hình B.2 – Kết quả dự đoán protein mới xuất ra dưới dạng file

Dưới đây là danh sách mô tả chi tiết

- (1): Release date là ngày kiểm thử (định dạng theo tháng/ngày/năm).
- (2): Algorithm type là loại mô hình đang được sử dụng (SVM hoặc K-NN).
- (3): Classification tree type là cây phân loại đang được sử dụng để dự đoán (SCOP hoặc CATH).
- (4): Mã protein trong SCOP hoặc CATH.
- (5): Mã chain của protein.
- (6): Mã protein.
- (7): Mã chain của protein.
- (8): Lớp dự đoán của mô hình, theo định dạng là Class.Fold nếu cây phân loại SCOP được sử dụng hoặc Classs.Architecture.Topology nếu cây phân loại CATH được sử dụng.
- (9): Tỉ lệ tương ứng của lớp dự đoán có thể xảy ra.
- (10): Mô tả lớp dự đoán.

## TÀI LIỆU THAM KHẢO

### Tiếng Việt

- [1] PGS. TS. Cao Đăng Nguyên, PGS. TS. Đỗ Quý Hải, Công Nghệ Protein, Khoa sinh học, Trường Đại học Huế, 2006.
- [2] Nguyễn Thanh Tùng, Ứng dụng mô hình Markov và cây quét định trong một số bài toán dự đoán, Luận Văn Thạc Sĩ Tin Học, ĐH Quốc Gia Tp. HCM, Trường Đại Học Khoa Học Tự Nhiên, 2004.
- [3] TS. Nguyễn Linh Giang, Nguyễn Mạnh Hiển, Phân loại văn bản tiếng Việt với bộ phân loại vector hỗ trợ SVM, Khoa công nghệ thông tin, Đại học Bách khoa Hà Nội, 2005.
- [4] Vũ Minh Thái, Lê Hoàng Hà, Xây Dựng Mô Hình Dự Đoán Cấu Trúc Bậc Cao Của Protein Bằng Phương Pháp Phân Lớp Dữ Liệu, khóa luận tốt nghiệp, ĐH Quốc Gia Tp. HCM, Trường Đại Học Công Nghệ Thông Tin, 2012.

### Tiếng Anh

- [5] PhD. Trần Nhân Dũng, Đỗ Tấn Khang, Nguyễn Thị Xuân Nguyên, Fabian Boes, Michael Knoll, Bio – Informatics, Can Tho University, 2008.
- [6] PhD. Betty Cheng, Bioinformatics in support of Molecular Medicine, Department of Medicine, Stanford University.
- [7] PhD. Amgad Madkour, Bioinformatics Research, Department of Computer Sciences, Purdue University.
- [8] CA Orengo, AD Michie, S Jones, DT Jones, MB Swindells and JM Thornton, CATH – a hierarchic classification of protein domain structures, Department of Biological Sciences, University of Warwick, 1997.

- [9] Alejandro Arbelaez, Youssef Hamadi, Michele Sebag, Building Portfolios for the Protein Structure Prediction Problem, Microsoft Research, Cambridge United Kingdom.
- [10] Jayanthi Sourirajan, Protein Structure Prediction, Computational Molecular Biology BIOC218, June 4, 2004.
- [11] Pualing & Corey. Configurations of Polypeptide Chain With Favored Orientations Around Single Bonds: Tow New Pleated Sheets. Canlifonia Institute of Teachnology, September 4, 1951.
- [12] Jiawei Han and Micheline Kamber, Data Mining: Concepts and Techniques, University of Illinois at Urbana – Champaign, 2006.
- [13] Zeyar Aung, Computational Analysis of 3D Protein Structures, National University of Singapore, 2006.
- [14] Jun Huan, Deepak Bandyopadhyay, Wei Wang, Jack Snoeyink, Jan Prins, and Alexander Tropsha, Comparing Graph Representations of Protein Structure for Mining Family – Specific Residue – Based Packing Motifs, Journal of Computational Biology, 2005.
- [15] Alan Fersht, Enzyme catalysis case study: Serine proteases, 1999.
- [16] Swetharg, Identifying the novel domain involved in human pathogenesis, Department of Bioinformatics, Affiliated to Thiruvalluvar University, India.
- [17] Gerd Anders, Matthias Nicola, Managing the Protein Data Bank with DB2 pure XML, Humboldt University.
- [18] David Clark, Molecular Biology, Southem Illinois Univesity, 2005.
- [19] PhD. Richard L.Gallo, Protein May Be Key to Psoriasis and Wound Care, San Diego Shool of Medicine, University of California.
- [20] wwPDB, Protein Data Bank Contents Guide: Atomic Coordinate Entry Format Description.

- [21] Hongyu Zhang, Protein Tertiay Structures: Prediction from Amino Acid Sequences, Univesity of Maryland, USA.
- [22] A. G. Murzin, S. E. Brenner, T. Hubbard, and C. Chothia, SCOP: a structural classification of protein database for the investigation of sequences and structures. *J. Mol. Biol.*, 247:536-540, 1995.
- [23] A. G. Murzin, S. E. Brenner, T. Hubbard, and C. Chothia, SCOP: a structural classification of protein database, 1997.
- [24] John C. Platt, Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines, Technical Report MSR-TR-98-14, 1998.
- [25] Yu-Feng Huang. Study of Mining Protein Structural Properties and its Application. National Taiwan University. December 11, 2007.
- [26] C. Cortes and V. Vapnik, Support vector networks, *Machine Learning*, 20, 273 – 297, 1995.
- [27] Nello Cristianini, Support Vector and Kernel Machines, BIOwulf Technologies, 2001.
- [28] J. P. Lewis, A Short Support Vector Machine Tutorial, U. Southern California, 2004.
- [29] Xindong Wu and Vipin Kumar, The Top Ten Algorithms in Data Mining, University of Minnesota, Deparment of Compute Science and Engineering, Minneapolis, Minnesota, U.S.A, 2009.
- [30] Helen M Berman, T. N. Bhat, Philip E. Bourne, Zukang Feng, Gary Gilliland, Helge Weissig, John Westbrook, The PDB and The Challenge of Structural Genomics, Research Collaboratory for Structural Bioinformatics.
- [31] Linderstrøm – Lang, Carlsberg Laboratory, The view of a postdoctoral fellow in 1954, Protein Science, Cambridge University Press, 1992.
- [32] Pualing & Corey. The pleated sheet, a new layer configuration of polypeptide chains. California Institute of Teachnology, March 31, 1951.

- [33] Pualing & Corey. The structure of feather rachins keratin. Canlifornia Institute of Teachnology, March 31, 1951.
- [34] Christian Cole, Jonathan D. Barber and Geoffrey J. Barton, The Jpred 3 secondary structure prediction server, University of Dundee, 2008.
- [35] Do Phuc, Nguyen Thi Kim Phung, Hoang Trong Nghia, Using Graph Spectral Analysis Comparison, IT@EDU2008 – HCM City, Vung Tau City.
- [36] Karsten M. Borgwardt, Cheng Soon Ong, Stefan Schonauer, S.V.N Vishwanathan, Alex J. Smola, Hans-Peter Kriegel, Protein function prediction via graph kernels, Bioinformatics, Vol. 00 no. 00 2005, Pages 1-9.
- [37] Morten Nielsen, Claus Lundgaard, Ole Lund and Thomas Nordahl Petersen, CPHmodels-3.0-remote homology modeling using structure-guided sequence profiles, The Technical University of Denmark, Denmark, 2010.
- [38] Cục y tế, cục phòng chống HIV/AIDS website -  
[http://www.vaac.gov.vn/Desktop.aspx/Hoi-dap/Hoi-dap/HIVAIDS\\_la\\_gi/](http://www.vaac.gov.vn/Desktop.aspx/Hoi-dap/Hoi-dap/HIVAIDS_la_gi/)
- [39] AForge.net website - <http://www.aforgenet.com/>