DỰ ĐOÁN PHÂN LOẠI CỦA ENZYME BẰNG CÁCH ÁP DỤNG KỸ THUẬT KHAI THÁC DỮ LIỆU ĐỒ THỊ

Phạm Quốc Đàm[1], Đỗ Phúc[2], Lê Thị Thanh Mai [3]
(1) Trường Đại học Tôn Đức Thắng, Tp.HCM
(2)Trường Đại học Công nghệ thông tin, ĐHQG-HCM, (3) ĐHQG-HCM
(Bài nhận ngày 25 tháng 06 năm 2007, hoàn chỉnh sửa chữa ngày 29 tháng 12 năm 2007)

TÓM TẮT: *Trong bài báo này, chúng tôi trình bày cách thức ứng dụng kỹ thuật khai thác dữ liệu để phân rã chuỗi amino acid cấu tạo nên enzyme - thuộc cùng một phân lớp enzyme đã được định danh - thành tập các đồ thị con phổ biến tối đại tương ứng. Các đồ thị con có thể có một đỉnh và cũng có thể có nhiều đỉnh. Khi cần dự đoán có một enzyme mới, thuộc phân lớp enzyme nào, ta chỉ cần phân rã chuỗi amino acid của enzyme đó, rồi so khớp với từng tập đồ thị con phổ biến tối đại, có trong cơ sở dữ liệu. Phân loại enzyme được dự đoán dựa trên phân loại có điểm số cao nhất sau khi so khớp. Việc thử nghiệm được triển khai dựa trên các phân lớp Oxidoreductase EC 1.1.1.1 và Hydrolase EC 3.1.1.3, đã cho kết quả tốt. Qua quá trình thử nghiệm, chúng tôi nhận thấy: khi mở rộng quy mô của tập học, nên chọn tất cả các enzyme đã được định danh. Mục đích của việc chọn lựa này là để tạo nên tập đồ thị con phổ biến tối đại có độ tin cậy cao.*

1. GIỚI THIỆU

Sự phát triển mạnh mẽ của công nghệ sinh học trong những năm gần đây đã tạo nên lượng dữ liệu rất lớn về enzyme (hơn 19000 enzymes). Trong khi đó, số lượng enzyme đã được định danh chính xác mới được khoảng 4006 enzymes. Vì vậy cần tìm kiếm phương pháp mới giúp dự đoán phân loại enzyme thoả các yêu cầu:

- Nhanh - Dễ sử dụng và
- Ít cần sự can thiệp của chuyên gia sinh học.

Khai thác dữ liệu đồ thị (Graph Mining) đang là một kỹ thuật mới, được dùng để phát hiện tri thức và đặc biệt thích hợp với dữ liệu có cấu trúc, vì có thể sử dụng đồ thị để mô tả.

Với enzyme, bộ 3 thành phần hoá học – cấu trúc – chức năng có quan hệ mật thiết với nhau. Vậy nếu có thể ứng dụng được Graph Mining để tìm được tập các đồ thị con chứa đặc trưng sinh học, việc phân loại enzyme có thể sẽ đạt hiệu quả hơn, hỗ trợ tốt cho chuyên gia sinh học trong quá trình định danh chính xác.

2. VẤN ĐỀ CẦN GIẢI QUYẾT

Graph Mining đang được nghiên cứu sử dụng nhiều trong lĩnh vực phân lớp văn bản. Đã có nhiều thành tựu được công bố trong các bài báo của các chuyên gia. Để có thể ứng dụng graph mining trong việc dự đoán phân loại enzyme, cần phải:

- Tìm cách biểu diễn enzyme dưới dạng đồ thị.
- Đề xuất phương pháp tìm tập đồ thị con chứa đặc trưng của enzyme bằng kỹ thuật graph mining sao cho đạt độ chính xác từ 70% trở lên.
- Đề xuất cách đánh giá và dự đoán phân loại enzyme.

3. CÁCH GIẢI QUYẾT

3.1. Để biểu diễn enzyme dưới dạng đồ thị, ta quy ước:

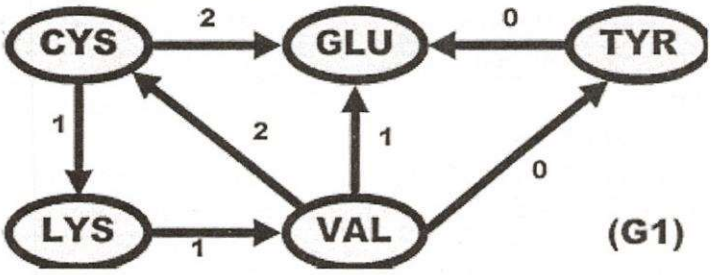
Mỗi amino acid được gọi là "Đỉnh" của đồ thị.

Sự đồng xuất hiện của hai đỉnh trong cấu trúc của enzyme sẽ có khả năng hình thành một "Cạnh" nối giữa hai đỉnh đó. Cạnh còn thể hiện khả năng xảy ra liên kết sinh – hoá giữa hai đỉnh.

Khoảng cách giữa 2 đỉnh được gọi là "nhãn cạnh" của đồ thị.

Ví dụ: với chuỗi amino acid của enzyme có dạng:

CYS xxx LYS GLU VAL TYR GLU CYS thì đồ thị biểu diễn sẽ là:

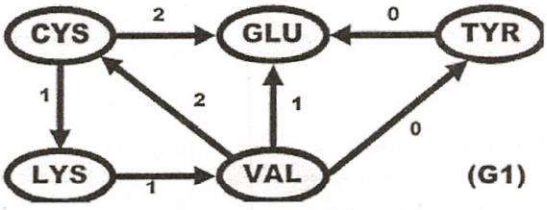


3.2. Phương pháp tìm tập đồ thị con phổ biến chứa đặc trưng:

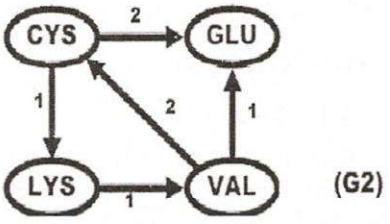
Sử dụng phương pháp AGM trên nền thuật toán Apriori để rút trích các đồ thị con có nghĩa. Các khái niệm:

Đồ thị con có nghĩa chính là đoạn amino acid (sub sequence), trích từ thành phần cấu trúc của enzyme (sequence residue) thoả tần suất cho trước.

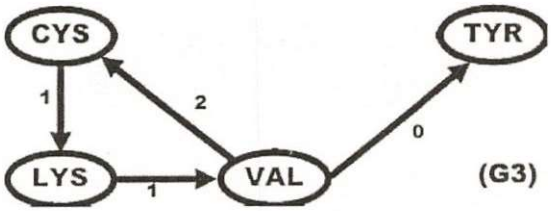
Ví dụ: ta có lớp enzyme EC chỉ chứa 4 enzyme có cấu trúc được biểu diễn dưới dạng đồ thị như sau:



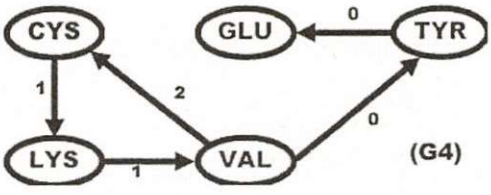
CYS xxx LYS GLU VAL TYR GLU CYS ...



CYS xxx LYS GLU VAL xxx GLU CYS

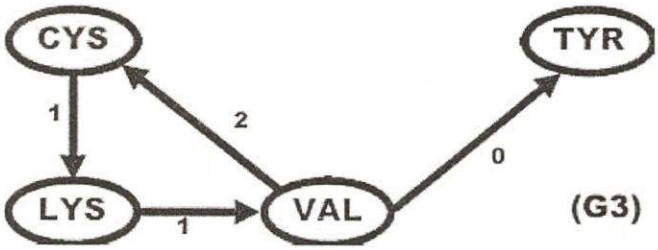


CYS xxx LYS xxx VAL TYR xxx CYS



CYS xxx LYS xxx VAL TYR GLU CYS

Với tần suất minsupp = 0.7 thì đồ thị con có chứa đặc trưng của lớp EC là:



Tập các đồ thị con có nghĩa ⇔ tập hợp các đoạn amino acid trích từ thành phần cấu trúc của loại enzyme.

Tập đồ thị tối đại lưu trữ đặc trưng ⇔ tập hợp các đoạn amino acid có đặc điểm không là đoạn con của đoạn có nhiều đỉnh hơn.

Cách giải quyết:

Cách phát sinh tập các đồ thị con có nghĩa:

Trích các đoạn amino acid từ thành phần cấu trúc của loại enzyme thoả tần xuất. Bắt đầu từ 1 amino acid (đồ thị có 1 đỉnh, bậc 1), tăng dần đến <n> amino acid (đồ thị có <n> đỉnh, bậc <n>).

Cách tìm tập đồ thị tối đại lưu trữ đặc trưng:

Đồ thị có nhiều đỉnh nhất (bậc n) là tập tối đại. Duyệt các đồ thị có n-1 đỉnh, đồ thị nào không chứa trong bất kỳ đồ thị n đỉnh nào, được gọi là đồ thị tối đại. Tương tự, duyệt cho đến đồ thị 1 đỉnh.

3.3. Đề xuất cách đánh giá và dự đoán phân loại enzyme

So khớp bằng cách kiểm tra số lượng đồ thị tối đại của từng lớp enzyme (trong tập học) có chứa trong thành phần cấu trúc của enzyme cần dự đoán hay không. Nếu có, tính điểm theo phương pháp tựa Naïve Bayes, cộng dồn điểm theo lớp enzyme so khớp.

Cách dự đoán: chọn phân lớp có điểm số cao nhất.

4. THỬ NGHIỆM

4.1. Tập mẫu học

Các PDB trong mỗi phân loại enzyme, download từ Protein Data Bank, được xếp thứ tự theo tên tập tin. Chọn các nhóm con có từ 5 tập tin trở lên, trích ra thành nhóm lớn hơn, có khoảng 50 PDB. Tiếp tục, chia các PDB trong nhóm lớn này thành 2 nhóm "học" và "test" theo tỷ lệ 8:2 áp dụng trên từng nhóm con PDB. Số lượng cụ thể như sau:

Bảng 4.1. Tập mẫu học

Loại enzyme	Số lượng trích		
	Học	Test	Cộng
Oxidoreductase EC 1.1.1.1 (Alcohol dehydrogenase)	40	10	50
Hydrolase EC 3.1.1.3 (Triacylglycerol lipase)	40	10	50

4.2. Các mẫu dự đoán phân loại

Việc dự đoán dựa trên kết quả học của 4.1. Có 3 trường hợp dự đoán phân loại sau:

4.2.1. Trường hợp 1: sử dụng các enzyme chưa được học (test) của 4.1

Bảng 4.2: Kết quả dự đoán phân loại enzyme của trường hợp 1.

Loại enzyme	Số mẫu học	Số mẫu test	Kết quả dự đoán	
			Đúng	Sai
Oxidoreductase EC 1.1.1.1	40	10	10 (100%)	0
Hydrolase EC 3.1.1.3	40	10	10 (100%)	0

Kết quả có thể được xem như là một bằng chứng cơ sở, chứng tỏ rằng có thể sử dụng Graph Mining vào việc tìm đặc trưng phục vụ cho việc phân lớp enzyme.

4.2.2. Trường hợp 2: sử dụng 100 enzyme chưa được học của 3 loại enzyme ở 4.1

Bảng 4.3: Loại enzyme và số lượng mẫu enzyme trích cho máy dự đoán

Loại enzyme	Số mẫu test	Ghi chú
Oxidoreductase EC 1.1.1.1	35	**10** của 4.2.1; **25** lấy mới.
Hydrolase EC 3.1.1.3	17	**10** của 4.2.1; **07** lấy mới.

Bảng 4.4: Kết quả dự đoán phân loại enzyme của trường hợp 2

Loại enzyme	Số mẫu học	Số mẫu test	Kết quả dự đoán	
			Đúng	Sai
Oxidoreductase EC 1.1.1.1	40	35	34 (97.14%)	1
Hydrolase EC 3.1.1.3	40	17	17 (100%)	0

4.2.3. Trường hợp 3: sử dụng 100 enzyme chưa được học của 3 loại enzyme ở 4.1, có thành phần đa dạng hơn so với 4.2.2

Bảng 4.5: Loại enzyme và số lượng enzyme trích cho máy dự đoán phân loại

Loại enzyme	Số mẫu test	Cộng	Ghi chú
Oxidoreductase EC 1.1.1.1	17	32	EC 1.1.3.38 chưa có mẫu học
EC 1.1.3.38	15	(=17+15)	
Hydrolase EC 3.1.1.3	17	34	EC 3.1.1.4 chưa có mẫu học
EC 3.1.1.4	17	(=17+17)	

Bảng 4.6: Kết quả dự đoán phân loại enzyme của trường hợp 3

Loại enzyme	Số mẫu học	Số mẫu test	Kết quả dự đoán	
			Đúng	Sai
Oxidoreductase EC 1.1.1.1	**40**	**17**	**17**	0

EC 1.1.3.38		**15**	(100%) **0** (0.0%)	15
Hydrolase EC 3.1.1.3	**40**	**17**	**17** (100%)	0
EC 3.1.1.4		**17**	**17** (100%)	0

4.2.4. Nguyên nhân các dự đoán sai:

So sánh với tập mẫu học, chúng tôi nhận thấy các enzyme có dự đoán sai là do trong tập mẫu học chưa có enzyme đồng dạng.

4.3. Tập đồ thị con phổ biến tối đại tìm được

Bảng 4.7.Số lượng đồ thị con tìm được.

Loại enzyme	Tần suất	Số đỉnh của đồ thị con	Số lượng đồ thị con tìm thấy	Số lượng đồ thị con tối đại
Oxidoreductase EC 1.1.1.1 (Alcohol dehydrogenase)				
	100%	1	18	1
	90%	2	3103	2697
	70%	3	332	330
	70%	4	1	1
			3454	**3029**
Oxidoreductase EC 1.1.1.1 (Alcohol dehydrogenase)				
	100%	1	19	0
	90%	2	12335	11160
	70%	3	1543	1245
	70%	4	237	139
	70%	5	69	45
	70%	6	14	12
	70%	7	1	1
			14218	**12602**

(*) Chúng tôi sử dụng 1 máy vi tính pentium 4, CPU 3GHz, 1 GB RAM để cài đặt chương trình chạy trên nền Windows XP Pro SP2.

4.4. Thời gian bình quân cho việc dự đoán phân loại của một enzyme: 17 phút.

5. HƯỚNG PHÁT TRIỂN VÀ CÁC THÁCH THỨC

Kết quả trên là cơ sở để tiếp tục triển khai nghiên cứu dự đoán phân loại enzyme trên tập dữ liệu có quy mô lớn hơn: thuộc nhiều phân lớp enzyme khác nhau, số lượng mẫu học nhiều hơn. Nếu có kết quả tốt, việc áp dụng graph mining sẽ là một hướng tiếp cận mới trong việc dự đoán phân loại enzyme.

Để có thể phát triển thành một công cụ hỗ trợ cho người dùng tại Việt Nam, chương trình cần đạt độ chính xác trên 90%; thời gian dự đoán cho mỗi mẫu cần nhanh hơn nữa và thời gian cập nhật một enzyme mới phải chấp nhận được. Khả năng đạt được các kết quả trên, thật sự là một thách thức.

PREDICTING THE SUB-CLASS OF ENZYME BY APPLYING GRAPH MINING BASED ON SEQUENCE STRUCTURE OF ENZYME

Pham Quoc Dam[(1)], Do Phuc[(2)], Le Thi Thanh Mai [(3)]
(1) Ton Duc Thang University, HoChiMinh city
(2)University of Information Technology, VNU-HCM
(3) VNU-HCM

ABSTRACT: *This article is a description of the way to apply graph mining technology to disaggregate the amino acid sequence of the enzyme - belonging to the same already named sub-class - into a set of respective maximal frequent subgraphs. The subgraphs can have one or many vertexes. When predicting the sub-class of a new enzyme, one just needs to disaggregate the amino acid sequence of that enzyme, then matches it with each maximal frequent subgraph in the data base. The predicted sub-class is based on the one with the highest scores after matching. The test developed on the sub-class of Oxidoreductase EC 1.2.1.1 and Hydrolase EC 3.1.1.3 gave good results. It left us with the remark that when enlarging the scale of learning set, all the named enzymes should be chosen. This aims to create a set of maximal frequent subgraphs with high reliability.*

TÀI LIỆU THAM KHẢO

[1]. Athony J. Cichoke, *The complete book of Enzyme Therapy*, Avery, (1999).
[2]. Artur M.Lesk, *Introduction to protein architecture*, University of Cambridge, (2001).
[3]. Manu Aery, Infosift – *Adapting Graph Mining Techniques For Document Classification*, The University of Texas at Arlington, December (2004).
[4]. Michael Knöll, Thai Ke Quan, *BioInformatic Training Course*, University of Stuttgart, (2006).
[5]. Nguyen Thanh Tung, Do Phuc, Tran Linh Thuoc, *Predicting protein secondary structure based on SCOP folds using HMM and DT,* Tạp chí Công nghệ Sinh học – Viện Khoa học - Công nghệ Việt Nam, số 2 – vol 4, trang 407-414, (2005).
[6]. Phuc Do, *Document Classification Using Graph Model, Frequent Subgraphs And Galois Lattice*, In PROC of IEEE RIVF 2006 conference, pages 173-176, (2006).
[7]. http://www.rcsb.org/pdb, RCSB PDB Protein Data Bank
[8]. http://www.chem.qmul.ac.uk/iubmb/enzyme/ *Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (NC-IUBMB)*
[9]. http://scop.mrc-lmb.cam.ac.uk/scop, *SCOP: Structural Classification of Protein.*
[10]. http://www.ncbi.nlm.nih.gov/, *NCBI: Nationel Center for Biotechnology Information.*
[11]. http://www.expasy.org/prosite, *Database of Protein Domains, families and functional sites.*