J. Sci. & Devel. 2015, Vol. 13, No. 2: 291-300 Tạp chí Khoa học và Phát triển 2015, tập 13, số 2: 291-300
www.vnua.edu.vn

PHÂN LOẠI GENE MÃ HÓA PROTEIN VẬN CHUYỂN SỬ DỤNG CÁC GENE HÀNG XÓM

Trần Vũ Hà*, Phạm Quang Dũng, Nguyễn Thị Thảo, Đoàn Thị Thu Hà

Khoa Công nghệ thông tin, Học viện Nông nghiệp Việt Nam

Email: tvha@vnua.edu.vn*

Ngày gửi bài: 06.10.2014 Ngày chấp nhận: 20.12.2014

TÓM TẮT

Cũng giống như sự đa dạng sinh học, trong tự nhiên có quá nhiều loại protein để chúng ta có thể miêu tả chức năng của chúng (anotate) bằng các thí nghiệm khoa học. Do đó các phương pháp để dự đoán chức năng của các protein trở nên cần thiết. Trong bài báo này chúng tôi đề xuất một phương pháp sử dụng dữ liệu sinh học để phân lớp các protein vận chuyển trên màng tế bào dựa vào cơ chất mà chúng vận chuyển. Dựa trên ý tưởng của các Operon, chúng tôi sử dụng dữ liệu biểu hiện gene và các GO terms của các gene hàng xóm để tạo dữ liệu đầu vào cho máy vector hỗ trợ. Để nhanh chóng thu được kết quả, chúng tôi tích hợp LIBSVM (A Library for Support Vector Machines) vào công cụ xử lý dữ liệu và sử dụng công cụ này để huấn luyện cũng như kiểm tra các bộ phân loại. Với công cụ này, người dùng có thể phân loại các protein vận chuyển và cả các loại protein khác; cho phép người dùng thêm dữ liệu của các sinh vật mới ngoài các sinh vật được sử dụng để thử nghiệm.

Từ khóa: Protein vận chuyển, gene hàng xóm, Gene Ontology.

Classifying Genes Encode Transmembrane Proteins Using Neighboring Genes

ABSTRACT

Like bio-diversity, there are too many proteins to experimentally annotate. Thus, methods for predicting the functions of proteins become necessary. In this article, we proposed a method that uses biological data to classify membrane transporters according to transported substrates. Motivated by the concept of Operons, our method used expression data and GO terms of neighboring genes to create input data for support vector machine. To rapidly obtain the result, we integrated LIBSVM in our tool then used this tool to train and test our classifiers. With this tool, users can classify membrane transporters and other kinds of proteins. This tool also allows users to add their desired organisms beside our tested ones.

Keywords: Gene Ontology, neighboring genes, transmembrane protein.

1. ĐẶT VẤN ĐỀ

Trong tự nhiên có rất nhiều loại protein khác nhau. Số lượng protein này một phần là do số lượng các loài sinh vật là rất lớn, một phần là do sự biến đổi của các phân tử trước khi hình thành nên protein hoàn chỉnh. Có hai sự biến đổi chính, thứ nhất là quá trình cắt/hợp của các chuỗi ribonucleic acid (RNA) sau khi chúng được phiên mã từ DNA (Black, 2003); thứ hai là sau quá trình dịch mã từ RNA thành chuỗi polypeptide, các chuỗi này tiếp tục trải qua các thay đổi khác (glycosylation hay phosphorylation) trước khi trở thành protein hoàn chỉnh. Thực tế này dẫn đến việc có rất nhiều protein chưa được giải thích bằng các thí nghiệm và vì vậy các phương pháp dự đoán chức năng của protein trở nên cần thiết.

Ngày nay có một vài cách tiếp cận khác nhau trong việc dự đoán chức năng của protein:

- Dự đoán chức năng dựa vào sự tương đồng của chuỗi polypeptide (homology-based): Đây là cách tiếp cận được sử dụng rộng rãi nhất trong việc dự đoán chức năng. Tuy nhiên, sự tương

đồng về trình tự chuỗi polypeptide của hai protein không đảm bảm rằng chúng có cùng chức năng ngay cả khi độ tương đồng của hai chuỗi là rất cao (Punta and Ofran, 2008).

- Sử dụng các motif (sequence motifs): Hiện nay có một số công cụ tính toán dành riêng cho việc xác định các motif như PRINT (Attwood et al., 1999), BLOCKS (Henikoff and S. Henikoff, 1996), PROSITE (Hofmann et al., 1999), InterPro (Apweiler et al., 2000), và ELM (Puntervoll et al., 2003). Các công cụ này thường cung cấp một thư viện lớn bao gồm các motif đã được thu thập bởi các chuyên gia, bởi các thuật toán hoặc bằng cách kết hợp cả hai phương pháp này (Punta and Ofran, 2008).

- Dự đoán dựa vào cấu trúc (structure-based) : Các protein tồn tại và hoạt động khi chúng có cấu trúc không gian 3 chiều (3D). Vì thế sự tương đồng về cấu trúc là một chỉ số tốt để xác định sự tương đồng về chức năng của hai hay nhiều protein (Sleator and Walsh, 2010; Whisstock and Lesk, 2003).

- Dự đoán dựa vào ngữ cảnh di truyền (genomic context-based): Các phương pháp này dựa vào các quan sát về hai hay nhiều protein có cùng sự xuất hiện hay vắng mặt trên các hệ gene khác nhau gần như chắc chắn có sự liên kết về mặt chức năng (Eisenberg et al., 2000; Sleator and Walsh, 2010).

- Dự đoán dựa vào mạng tương tác protein (protein-protein interaction networks): Trong các mạng này, các nút mạng là các gene/protein và được liên kết với nhau bởi các cạnh thể hiện sự chia sẻ chức năng giữa chúng (Sharan et al., 2007).

Trong mỗi cách tiếp cận, sự tương đồng trong cấu trúc hay sự tương đồng về tương tác được xem như các bằng chứng về sự tương đồng chức năng. Mỗi cách tiếp cận có ưu điểm và nhược điểm riêng. Ở đây, chúng tôi kết hợp dự đoán dựa vào ngữ cảnh di truyền với Gene Ontology (GO) để tạo ra một phương pháp dự đoán mới. Lý do mà chúng tôi chọn phương pháp dự đoán dựa vào ngữ cảnh di truyền là vì phương thức này sử dụng vị trí và sự đồng biểu hiện của các gene và đây cũng là ý tưởng của Operon và các đặc tính của nó. Được đề cập lần đầu tiên vào năm 1960 bởi Jacob và các cộng sự, một operon là một nhóm các gene mà sự biểu hiện của chúng được điều khiển bởi một promoter duy nhất (Jacob et al., 1960). Vì được điều khiển bởi một đơn vị (promoter) nên các gene trong một operon được biểu hiện cùng nhau hoặc không gene nào được biểu hiện. Do đó chúng cũng thường có chức năng tương tự nhau. Thông thường, các operon tồn tại trong các sinh vật nguyên thủy (prokaryote) nhưng trong một số ít các trường hợp chúng cũng được tìm thấy trong các sinh vật nhân điển hình (eukaryote). Trong khi các phương pháp dự đoán chức năng protein dựa vào ngữ cảnh di truyền được ủng hộ bởi các operon trong các sinh vật nguyên thủy thì mục tiêu của Gene Ontology Consortium là tạo nên một bộ từ vựng có thể sử dụng cho mọi sinh vật nhân điển hình (Ashburner et al., 2000). Bằng cách kết hợp hai kỹ thuật này, chúng tôi dự định tạo ra một kỹ thuật có thể áp dụng cho cả sinh vật nguyên thủy và sinh vật nhân điển hình.

2. VẬT LIỆU VÀ PHƯƠNG PHÁP

2.1. Vật liệu nghiên cứu

Trong nghiên cứu này chúng tôi lựa chọn hai nhóm là protein vận chuyển amino acid và protein vận chuyển đường (đường). Cụ thể là 27 gene mã hóa protein vận chuyển amino acid (AVT6, AVT3, GNP1, AVT4, GAP1, AVT1, VBA3, VBA1, VBA2, BAP3, MMP1, AGC1, DIP5, TAT1, TAT2, HIP1, PUT4, ODC1, CAN1, ODC2, MUP3, ATG22, ALP1, SAM3, AGP3, SSY1, LYP1) và 24 gene mã hóa protein vận chuyển đường (GIT1, MAL31, HXT1, MAL11, VRG4, H6XT2, HXT3, GAL2, ITR1, ITR2, STL1, SNF3, HXT17, RGT2, HXT15, HXT16, MPH3, HXT13, HXT14, HXT8, MPH2, HXT5, HXT7, HXT) của *Saccharomyces cerevisiae*. Với *Escherichia coli,* chúng tôi sử dụng 30 gene mã hóa protein vận chuyển amino acid (MmuP,

metN, TdcC, LysP, HisP, LivG, CycA, YgjU, GltL, TyrP, GlnQ, rhtB, rhtC, BrnQ, PotE, YecC, TauB, YbiF, GltS, AroP, GltP, ArtP, CadB, PutP, YjdE, PheP, TnaB, ProP, SdaC, Mtr) và 27 gene mã hóa protein vận chuyển đường (GalP, SetA, XylE, NanT, MalK, XylG, MtlA, MelB, alsA, UhpT, LacY, ManY, AscF, setB, TreB, PtsG, SotB, CelB, AraE, AraG, GlvC, RhaT, NagE, FruB, BglF, RbsA, FucP) (Barghash and Helms, 2013).

Các gene hàng xóm của *Escherichia coli* được tải từ EcoCyc (http://ecocyc.org/ download.shtml) và của *Saccharomyces cerevisiae* được tải từ UCSC (genome-mysql.cse.ucsc.edu).

Dữ liệu biểu hiện gene của *Escherichia coli* và *Saccharomyces cerevisiae* được tải từ GEO. Với *Escherichia coli* chúng tôi sử dụng DataSet Record GDS2768 (Domka et al., 2007), còn đối với *Saccharomyces cerevisiae* chúng tôi sử dụng DataSet Record GDS9 (Brem et al., 2002).

Để tìm GO term - thành phần cơ bản của GO, mỗi thuật ngữ (term) mô tả một thuộc tính của gene hoặc protein - cho các gene cần thông qua hai bước chính. Đầu tiên, chúng tôi tìm UniProt ID cho gene symbol từ http://www.kegg.jp. Tiếp theo, tìm GO term cho tất cả các UniProt ID từ http://uniprot.org.

2.2. Phương pháp nghiên cứu

Các dữ liệu sau khi xử lý được đưa vào phân loại, sử dụng phương pháp vector hỗ trợ (support vector machine - SVM) thông qua công cụ nổi tiếng có tên LIBSVM (Chang and Lin, 2011).

3. KẾT QUẢ VÀ THẢO LUẬN

3.1. Phân loại protein vận chuyển sử dụng gene hàng xóm

Một cách đơn giản, phương pháp phân loại protein mà chúng tôi đề xuất được trình bày trong hình 1.

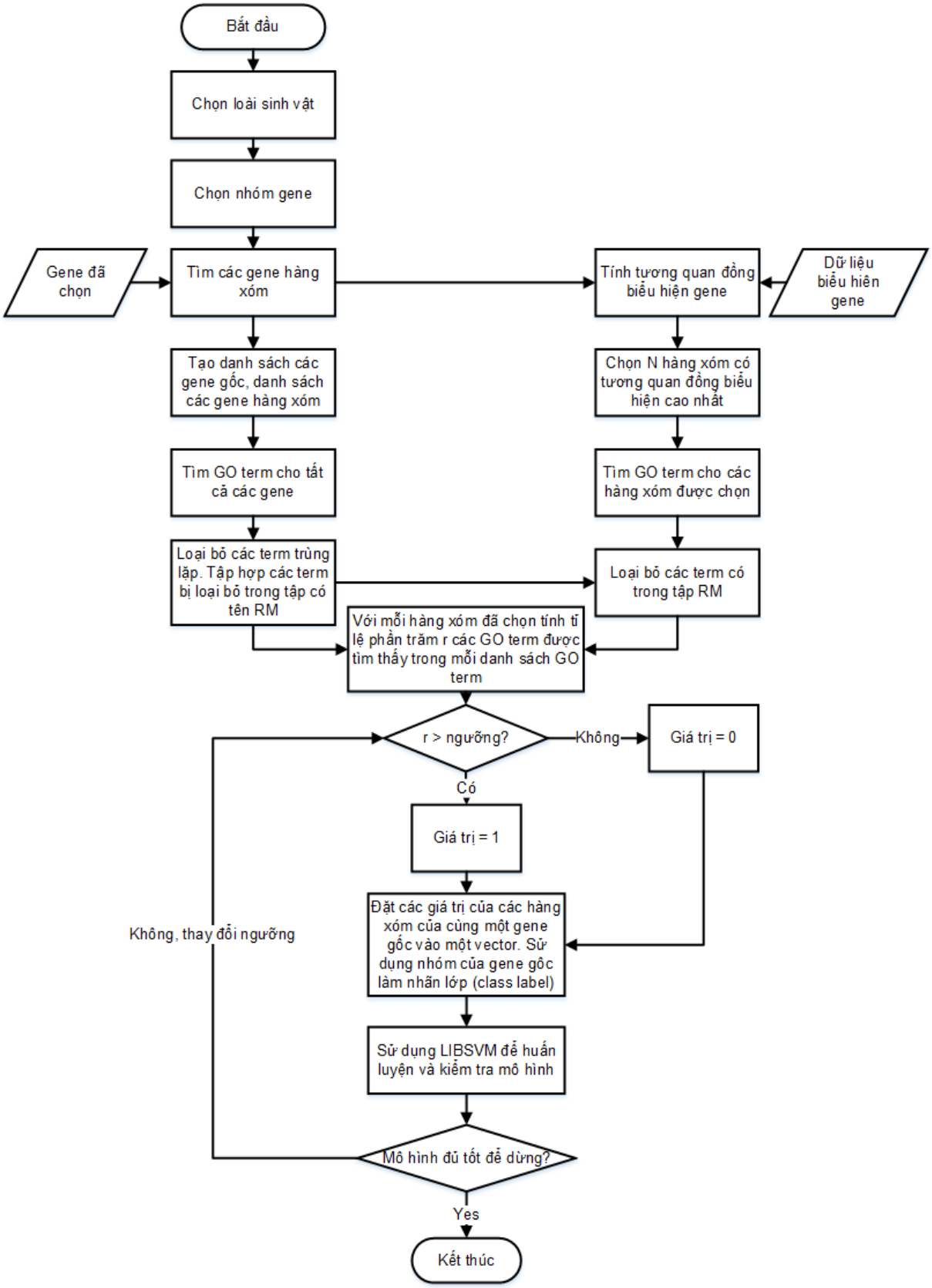
3.2. Công cụ xử lý dữ liệu

Để phát triển công cụ xử lý số liệu, chúng tôi lựa chọn ngôn ngữ lập trình Java. Đây là ngôn ngữ lập trình để phát triển phần mềm cho nhiều loại thiết bị (máy tính để bàn, máy chủ, thiết bị di động và các thiết bị nhúng). Để tạo một lượng lớn các ứng dụng cho thiết bị di động, máy tính cá nhân và các máy chủ, Java được cung cấp theo ba ấn bản (editions): Java Standard Edition (Java SE), Java Enterprise Edition (Java EE), Java Micro Edition (Java ME). Ngày nay, Java trở nên phổ biến nhờ những đặc điểm đáng chú ý như: thuần hướng đối tượng, phân tán, đa luồng và có thể chạy trên nhiều nền tảng (platform) khác nhau mà không cần sửa đổi mã nguồn chương trình.

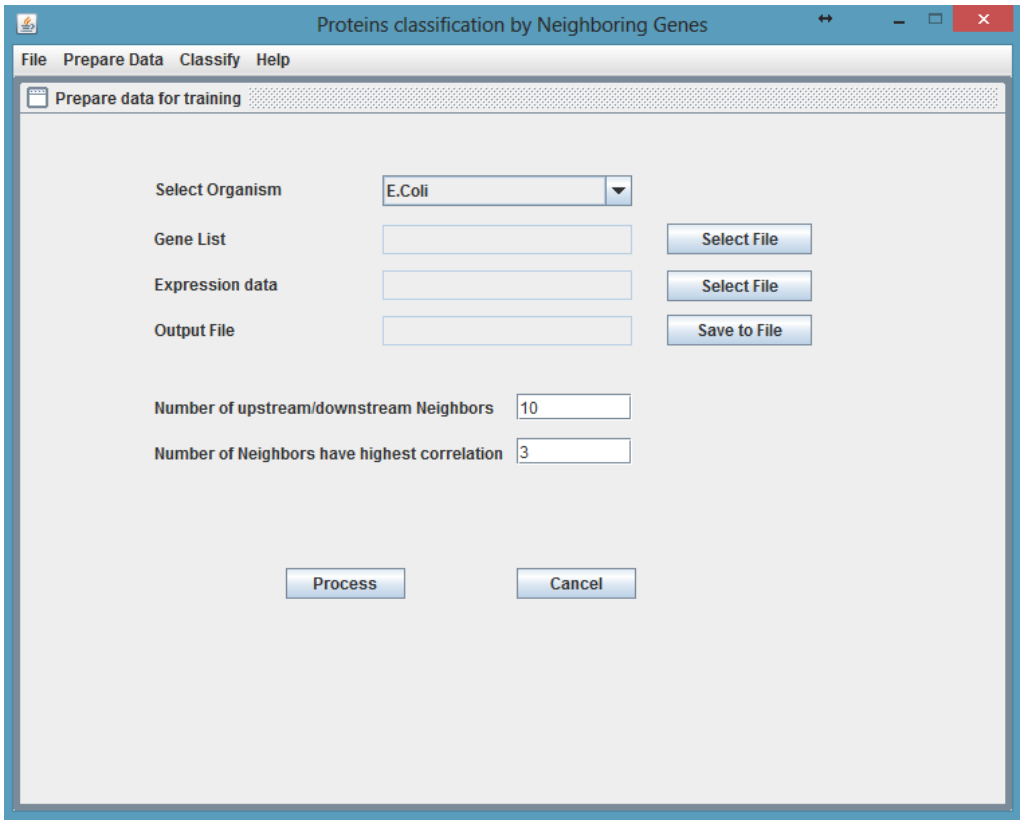
Trong đề tài này chúng tôi sử dụng Java SE để phát triển công cụ xử lý dữ liệu và phân lớp. Có nhiều phiên bản Java khác nhau và chúng tôi lựa chọn phiên bản 7 của Java SE. Mỗi phiên bản Java SE được phát hành cùng với một Java Development Kit (JDK). Với Java SE 7, Java Development Kit được gọi là JDK 1.7. JDK bao gồm các chương trình được sử dụng để phát triển và kiểm thử phần mềm, tuy nhiên các chương trình này thường yêu cầu người dùng tương tác qua dòng lệnh. Để thuận tiện hơn người dùng có thể sử dụng các công cụ phát triển với giao diện đồ họa (graphical user interface - GUI) như NetBeans, Eclipse hay JCreator (chúng tôi chọn sử dụng Eclipse). Các công cụ này cung cấp một môi trường phát triển tích hợp (integrated development environment - IDE) cho phép soạn thảo mã nguồn, dịch chương trình, gỡ lỗi trong cùng một GUI.

Công cụ phần mềm của chúng tôi có 4 chức năng chính (Hình 2, 4):

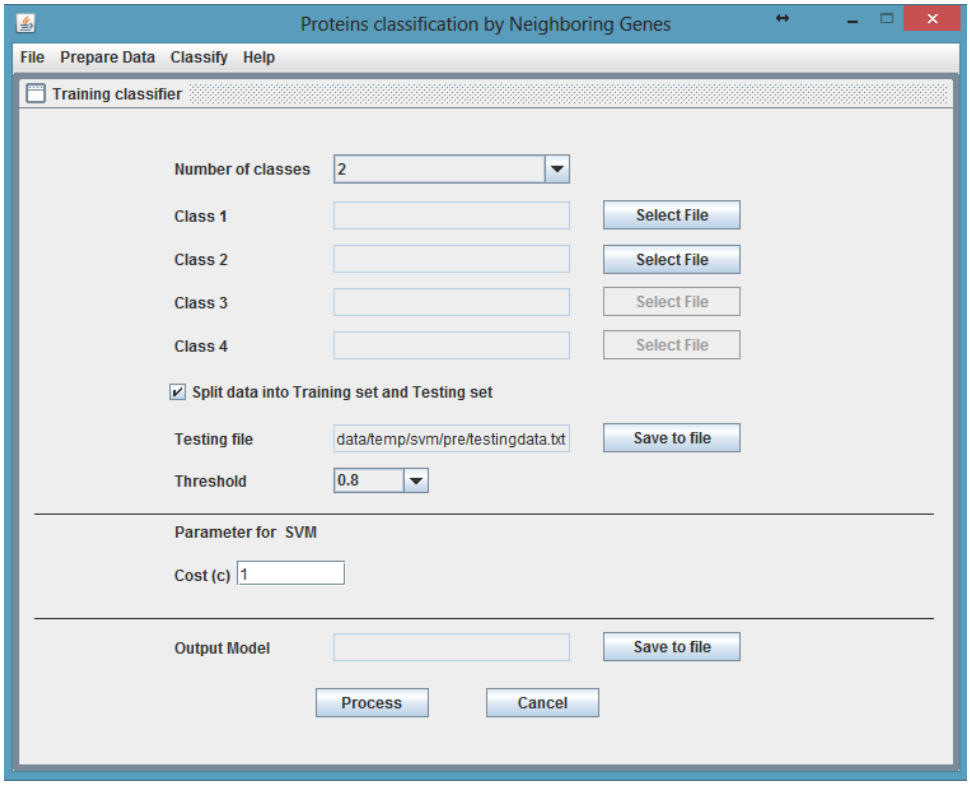
- Chuẩn bị dữ liệu cho quá trình huấn luyện
- Huấn luyện mô hình (bộ phân loại)
- Chuẩn bị dữ liệu để kiểm tra mô hình
- Kiểm tra mô hình



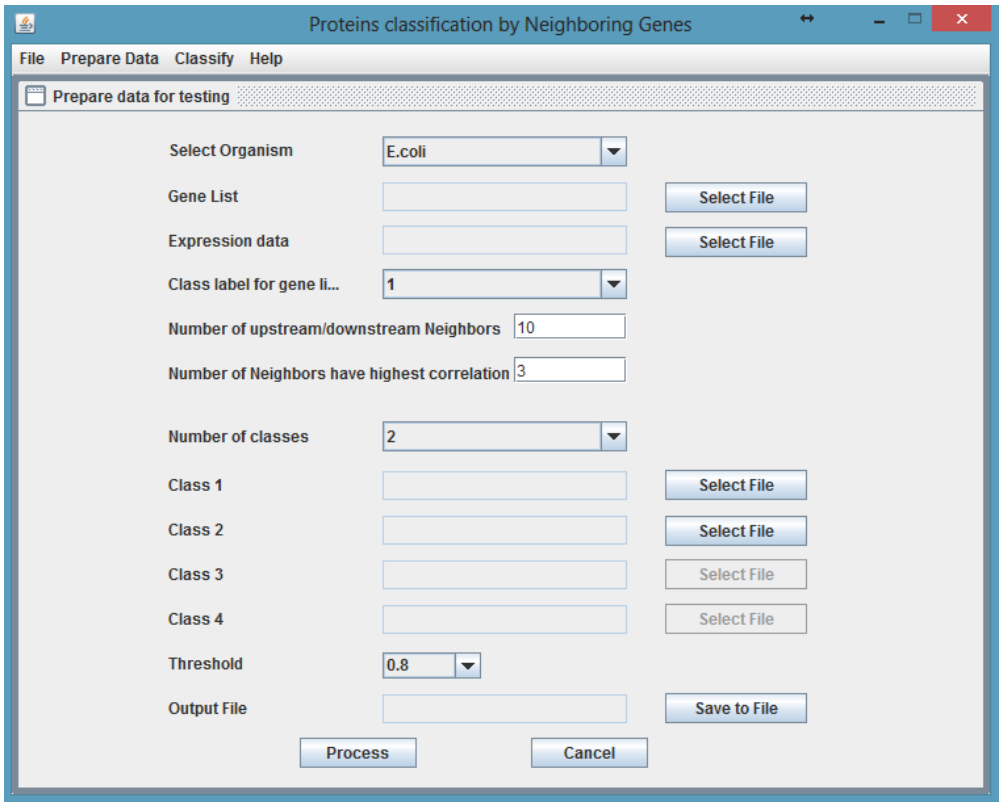
Hình 1. Phương pháp phân loại protein sử dụng gene hàng xóm



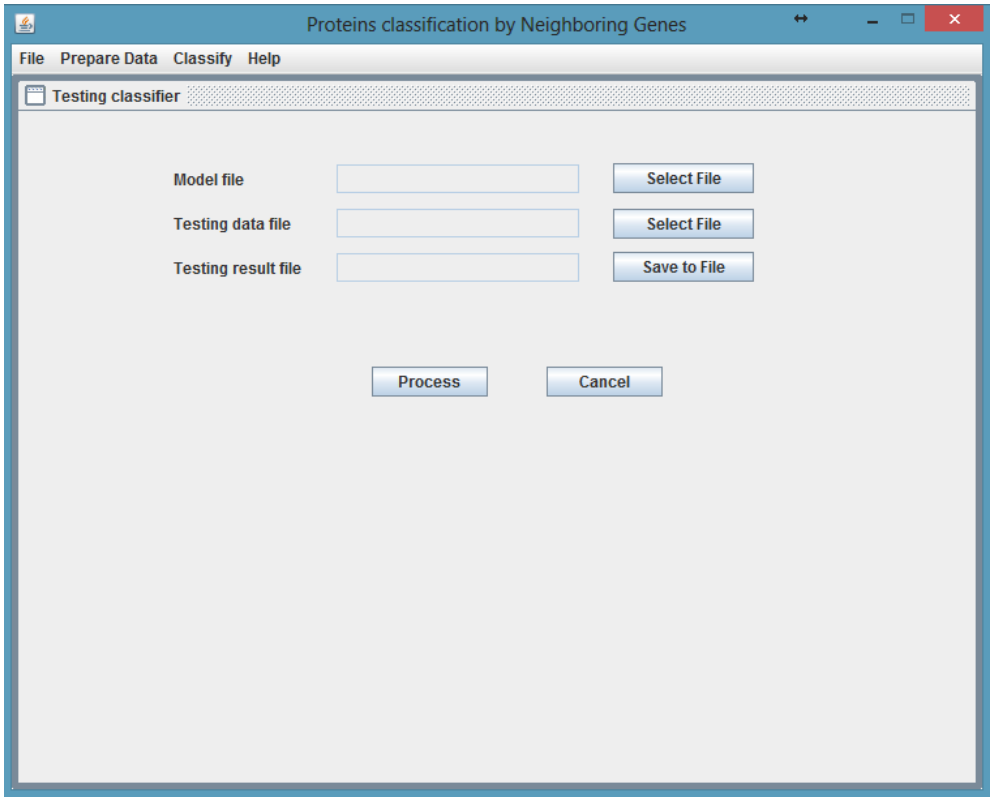
Hình 2. Form chuẩn bị dữ liệu huấn luyện mô hình



Hình 3. Form huấn luyện mô hình



Hình 4. Form chuẩn bị dữ liệu kiểm tra mô hình



Hình 5. Form kiểm tra mô hình

3.3. Kết quả phân loại

Với *Escherichia coli*, chúng tôi thu thập 30 gene mã hóa protein vận chuyển amino acid và 27 gene mã hóa protein vận chuyển đường, tuy nhiên thực tế chỉ có 26 gene vận chuyển amino acid và 24 gene vận chuyển đường là có thể sử dụng cho việc huấn luyện hoặc kiểm tra các bộ phân loại (với các gene khác, chúng tôi không thể tìm được gene hàng xóm hoặc dữ liệu về biểu hiện gene không có sẵn). Với mỗi gene chúng tôi tìm 10 hàng xóm nằm bên phải và 10 hàng xóm nằm bên trái, sau đó lựa chọn 3 hàng xóm có mức độ đồng biểu hiện cao nhất với gene trung tâm. Sau đó chúng tôi chọn ngưỡng cho tỉ lệ phần trăm r là 0,8. Sau đó dữ liệu cho SVM được tạo và được trình bày trong bảng 1và bảng 2. Trong cả hai bảng, List 1 đại diện cho danh sách GO term của tất cả các gene mã hóa protein vận chuyển amino acid, List 2 đại diện cho danh sách GO term của tất cả các hàng xóm của các gene mã hóa protein vận chuyển amino acid, List 3 đại diện cho danh sách GO term của tất cả các gene mã hóa protein vận chuyển đường, List 4 đại diện cho danh sách GO term của tất cả các hàng xóm của các gene mã hóa protein vận chuyển đường (đường).

Bảng 1. Dữ liệu cho SVM được tạo bởi các gene vận chuyển amino acid của *Escherichia coli*

Class label	Neighbors 1				Neighbors 2				Neighbors 3			
	List 1	List 2	List 3	List 4	List 1	List 2	List 3	List 4	List 1	List 2	List 3	List 4
1	0	1	0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	1	0	0
1	0	1	0	0	0	1	0	0	0	1	0	0
1	0	0	0	0	0	0	0	0	0	0	0	0
1	1	1	0	0	1	1	0	0	0	1	0	0
1	0	0	0	0	1	1	0	0	1	1	0	0
1	0	0	0	0	0	0	0	0	0	0	0	0
1	1	1	0	0	0	0	0	0	1	1	0	0
1	0	1	0	0	0	0	0	0	0	0	0	0
1	1	1	0	0	1	1	0	0	0	0	0	0
1	0	0	0	0	0	1	0	0	0	1	0	0
1	0	1	0	0	1	1	0	0	0	1	0	0
1	0	1	0	0	0	0	0	0	0	1	0	0
1	0	1	0	0	0	1	0	0	0	1	0	0
1	0	0	0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	1	0	0	0	0	0	0
1	0	0	0	0	0	1	0	0	0	1	0	0
1	0	1	0	0	0	1	0	0	0	1	0	0
1	0	0	0	0	0	0	0	0	0	1	0	0
1	0	1	0	0	0	0	0	0	0	1	0	0
1	0	1	0	0	0	1	0	0	0	1	0	0
1	0	0	0	0	0	1	0	0	0	0	0	0
1	0	1	0	0	0	0	0	0	0	1	0	0
1	0	0	0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0	0	0
1	0	1	0	0	0	0	0	0	0	0	0	0

Bảng 2. Dữ liệu cho SVM được tạo bởi các gene vận chuyển đường của *Escherichia coli*

Class label	Neighbors 1				Neighbors 2				Neighbors 3			
	List 1	List 2	List 3	List 4	List 1	List 2	List 3	List 4	List 1	List 2	List 3	List 4
2	0	0	0	1	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	1	0	0	0	1
2	0	0	0	1	0	0	0	0	0	0	0	1
2	0	0	0	0	0	0	0	0	0	0	0	0
2	0	0	0	1	0	0	0	1	0	0	0	0
2	0	0	0	0	0	0	0	1	0	0	0	1
2	0	0	0	1	0	0	0	1	0	0	0	1
2	0	0	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	1	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0	0	0
2	0	0	0	1	0	0	0	0	0	0	0	1
2	0	0	1	1	0	0	1	1	0	0	0	1
2	0	0	0	1	0	0	0	1	0	0	0	1
2	0	0	0	1	0	0	0	1	0	0	0	1
2	0	0	0	0	0	0	0	0	0	0	0	0
2	0	0	0	1	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	1	0	0	0	1
2	0	0	0	0	0	0	0	0	0	0	0	0
2	0	0	0	1	0	0	1	1	0	0	0	1
2	0	0	0	1	0	0	0	1	0	0	0	0
2	0	0	1	1	0	0	0	1	0	0	0	0
2	0	0	0	0	0	0	1	1	0	0	0	0
2	0	0	0	0	0	0	0	1	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0	0	0

Lựa chọn ngẫu nhiên 14 trong số 26 gene vận chuyển amino acid và 13 trong số 24 gene vận chuyển đường để sử dụng vào huấn luyện mô hình, số còn lại sử dụng để kiểm tra mô hình. Quá trình này được lặp lại 10 lần. Sau khi huấn luyện chúng tôi có các bộ phân loại với độ chính xác trung bình khi phân loại là 78,26% (18/23 gene của bộ dữ liệu kiểm tra được phân loại chính xác).

Với *Saccharomyces cerevisiae,* thực hiện tương tự như với *Escherichia coli.* Dữ liệu cho SVM của *Saccharomyces cerevisiae* được chỉ tra trong bảng 3 và bảng 4.

Lựa chọn ngẫu nhiên 12 gene trong nhóm amino acid và 6 gene trong nhóm đường để huấn luyện bộ phân loại. Số gene còn lại được sử dụng để kiểm tra bộ phân loại. Qua 10 lần như vậy các bộ phân loại có độ chính xác trung bình là 85,71%, tương đương với 12/14 gene được phân lớp chính xác.

Giờ hãy xem chi tiết hơn 4 bảng dữ liệu (Bảng 1- 4). Dễ nhận thấy các gene trong nhóm amino acid chứa giá trị 0 trong các cột List 3 và List 4 trong khi có rất nhiều giá trị 1 trong cột List 2. Với nhóm đường, các giá trị trong cột List 1 và List 2 đều bằng 0 trong khi có rất nhiều giá trị 1 ở cột List 4. Nguyên nhân của việc này là GO term của các hàng xóm được lựa chọn cho nhóm amino acid đều đã được bao gồm trong List 2 (danh sách GO term của tất cả các gene

Bảng 3. Dữ liệu cho SVM được tạo bởi các gene vận chuyển amino acid của *Saccharomyces cerevisiae*

Class Label	Neighbor 1				Neighbor 2				Neighbor 3			
	List 1	List 2	List 3	List 4	List 1	List 2	List 3	List 4	List 1	List 2	List 3	List 4
1	0	1	0	0	0	1	0	0	0	1	0	0
1	0	1	0	0								
1	0	1	0	0	0	1	0	0	0	0	0	0
1	0	1	0	0	0	1	0	0	0	1	0	0
1	0	0	0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0	0	0
1	0	1	0	0	0	1	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	1	0	0
1	0	1	0	0	0	1	0	0	0	1	0	0
1	0	0	0	0	0	1	0	0	0	1	0	0
1	0	1	0	0	0	0	0	0	0	1	0	0
1	0	1	0	0								
1	0	1	0	0	0	0	0	0	0	0	0	0
1	0	1	0	0	0	1	0	0	0	1	0	0
1	0	0	0	0	0	1	0	0	0	0	0	0
1	0	1	0	0	0	1	0	0	0	1	0	0
1	0	0	0	0	0	1	0	0	0	1	0	0
1	0	1	0	0	0	1	0	0	0	1	0	0
1	0	1	0	0	0	1	0	0	0	1	0	0

Bảng 4. Dữ liệu cho SVM được tạo bởi các gene vận chuyển đường của *Saccharomyces cerevisiae*

Class Label	Neighbor 1				Neighbor 2				Neighbor 3			
	List 1	List 2	List 3	List 4	List 1	List 2	List 3	List 4	List 1	List 2	List 3	List 4
2	0	0	0	0	0	0	0	0	0	0	0	0
2	0	0	1	0	0	0	0	1	0	0	0	0
2	0	0	0	1	0	0	0	0	0	0	0	1
2	0	0	0	1	0	0	1	1	0	0	0	0
2	0	0	0	1	0	0	0	1	0	0	0	1
2	0	0	1	1	0	0	0	0	0	0	0	1
2	0	0	0	1	0	0	0	0	0	0	0	1
2	0	0	0	1	0	0	0	1				
2	0	0	0	1	0	0	0	1	0	0	0	0
2	0	0	0	1	0	0	0	1	0	0	0	0
2	0	0	1	1	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	1	1	0	0	0	0

hàng xóm của nhóm amino acid) và tất cả các term trùng lặp của (List 2 và List 3) với (List 2 và List 4) đã bị loại bỏ. Với các giá trị 0 trong cột List 2 của nhóm amino acid, tất cả các GO terms của các hàng xóm được lựa chọn đã bị loại bỏ vì các term này cùng xuất hiện trong List 2 và List 3 hoặc cùng xuất hiện trong List 2 và List 4. Việc giải thích cho các gene trong nhóm đường cũng hoàn toàn tương tự như các gene trong nhóm amino acid. Chính đặc điểm này của

các bảng dữ liệu đã cho thấy các gene hàng xóm của các gene trong nhóm amino acid và các gene hàng xóm của các gene trong nhóm đường khác nhau về chức năng và nó cũng giúp chúng ta thấy lý do tại sao độ chính xác của các bộ phân loại lại cao như vậy.

4. KẾT LUẬN

Bài báo này đã trình bày một phương thức đơn giản để phân loại các protein vận chuyển theo cơ chất tương ứng có sử dụng dữ liệu biểu hiện gene và GO term của các gene hàng xóm bằng kỹ thuật phân loại SVM. Chúng tôi đã kiểm tra phương pháp của mình với các gene mã hóa protein vận chuyển amino acid và đường của 2 sinh vật là *Escherichia coli* và *Saccharomyces cerevisiae*.

Một công cụ phân loại sử dụng ngôn ngữ lập trình Java đã được phát triển để người dùng có thể thu được kết quả phân loại dễ dàng và thuận tiện hơn. Công cụ này không giới hạn trong việc phân lớp các gene mã hóa protein vận chuyển, người dùng có thể dùng nó để phân lớp các gene thuộc các metabolic pathways khác nhau hoặc các gene mã hóa các nhóm protein khác nhau. Công cụ này cũng không bị giới hạn trong các sinh vật như *Escherichia coli* hay *Saccharomyces cerevisiae*, người dùng có thể phân lớp các gene từ những sinh vật khác nữa.

TÀI LIỆU THAM KHẢO

Apweiler, R.; T. K. Attwood; A. Bairoch; A. Bateman; E. Birney; M. Biswas; P. Bucher; L. Cerutti; F. Corpet; M. D. Croning, et al. (2000). "Interpro--an Integrated Documentation Resource for Protein Families, Domains and Functional Sites." Bioinformatics, 16(12): 1145-50.

Ashburner, M.; C. A. Ball; J. A. Blake; D. Botstein; H. Butler; J. M. Cherry; A. P. Davis; K. Dolinski; S. S. Dwight; J. T. Eppig, et al. (2000). "Gene Ontology: Tool for the Unification of Biology. The Gene Ontology Consortium." Nat Genet, 25(1): 25-9.

Attwood, T. K.; D. R. Flower; A. P. Lewis; J. E. Mabey; S. R. Morgan; P. Scordis; J. N. Selley and W. Wright. (1999). "Prints Prepares for the New Millennium." Nucleic Acids Res, 27(1): 220-5.

Barghash, A. and V. Helms (2013). "Transferring Functional Annotations of Membrane Transporters on the Basis of Sequence Similarity and Sequence Motifs." BMC Bioinformatics, 14: 343.

Black, D. L. (2003). "Mechanisms of Alternative Pre-Messenger Rna Splicing." Annu Rev Biochem., 72: 291-336.

Brem, R. B.; G. Yvert; R. Clinton and L. Kruglyak (2002). "Genetic Dissection of Transcriptional Regulation in Budding Yeast." Science, 296(5568). 752-5.

Chang, C. C. and C. J. Lin (2011). "Libsvm: A Library for Support Vector Machines." Acm Transactions on Intelligent Systems and Technology, 2(3)1-27.

Domka, J.; J. Lee; T. Bansal and T. K. Wood (2007). "Temporal Gene-Expression in Escherichia Coli K-12 Biofilms." Environ Microbiol., 9(2): 332-46.

Eisenberg, D.; E. M. Marcotte; I. Xenarios and T. O. Yeates (2000). "Protein Function in the Post-Genomic Era." Nature, 405(6788): 823-6.

Henikoff, J. G. and S. Henikoff (1996). "Blocks Database and Its Applications." Methods Enzymol, 266: 88-105.

Hofmann, K.; P. Bucher; L. Falquet and A. Bairoch (1999). "The Prosite Database, Its Status in 1999." Nucleic Acids Res, 27(1): 215-9.

Jacob, F.; D. Perrin; C. Sanchez and J. Monod (1960). "[Operon: A Group of Genes with the Expression Coordinated by an Operator]." C R Hebd Seances Acad Sci., 250: 1727-9.

Punta, M. and Y. Ofran. 2008. "The Rough Guide to in Silico Function Prediction, or How to Use Sequence and Structure Information to Predict Protein Function." PLoS Comput Biol., 4(10), e1000160.

Puntervoll, P.; R. Linding; C. Gemund; S. Chabanis-Davidson; M. Mattingsdal; S. Cameron; D. M. Martin; G. Ausiello; B. Brannetti; A. Costantini, et al. (2003). "Elm Server: A New Resource for Investigating Short Functional Sites in Modular Eukaryotic Proteins. "Nucleic Acids Res., 31(13): 3625-30.

Sharan, R.; I. Ulitsky and R. Shamir (2007). "Network-Based Prediction of Protein Function." Mol Syst Biol., 3: 88.

Sleator, R. D. and P. Walsh (2010). "An Overview of in Silico Protein Function Prediction." Arch Microbiol., 192(3): 151-5.

Whisstock, J. C. and A. M. Lesk (2003). "Prediction of Protein Function from Protein Sequence and Structure." Q Rev Biophys., 36(3): 307-40.