ĐẠI HỌC QUỐC GIA HÀ NỘI
ĐẠI HỌC KHOA HỌC TỰ NHIÊN
Khoa Toán - Cơ - Tin

MỘT CÁCH TIẾP CẬN TÍNH TOÁN ÁP DỤNG SIMULATED ANNEALING ĐỂ NGHIÊN CỨU TÍNH ỔN ĐỊNH CỦA MẠNG TƯƠNG TÁC PROTEIN TRONG UNG THƯ VÀ RỐI LOẠN THẦN KINH.

Nhóm Bảo Vệ:
Tạ Văn Nhân
Nguyễn Thành Trung

Hà Nội, 7-12-2019

Mục lục

Tóm Lược Mạng phân tử cung cấp một công cụ mạnh mẽ cho việc nghiên cứu hệ thống sinh học, một vài nghiên cứu cụ thể đã phát hiện ra những thay đổi của cấu trúc mạng liên quan đến các trạng thái của bệnh. Có những bệnh không chỉ thay đổi cấu trúc của mạng mà cả sự ổn định của nó. Một phương pháp luận mới để đánh giá sự ổn định của mạng là áp dụng của thuật toán Deterministic Simulated Annealing cổ điển để làm việc với các trạng thái rời rạc. Giá trị năng lượng điều chỉnh được sử dụng để so sánh sự ổn định của mạng trong các trạng thái đối chứng và trạng thái bị bệnh. Những kết quả chỉ ra rằng mạng ung thư ổn định kém hơn mạng Alzheimer (AD). Những kết quả này có thể được giải thích thông qua các quan sát trước đó về sự đối nghịch của ung thư và AD, ví dụ các bệnh nhân AD có ít nguy cơ bị ung thư hơn.

1 Giới thiệu

Rối loạn thần kinh và ung thư là hai bệnh được đặc biệt quan tâm hiện nay. Bằng chứng dịch tễ học cho thấy bệnh nhân mắc một số rối loạn thần kinh, bao gồm cả những người mắc bệnh tâm thần phân liệt (SCZ) và bệnh Alzheimer(AD), có xu hướng ít mắc một số dạng ung thư hơn(BehDR et al. 2009, 2012; Tabarés-Seisdedos và Rubenstein 2013; Tabarés-Seisdedos et al.2011). Do đó, một phân tích tổng hợp biểu hiện gen được thực hiện để phát hiện các cơ chế phân tử có thể gây ra tình trạng đối nghịch như: xác định gen và những đường biểu hiện khác nhau trong rối loạn thần kinh và một số loại ung thư (Ibáñez et al. 2014).

Theo luận thuyết trung tâm (center dogma), DNA là trung tâm của di truyền, DNA tạo ra RNA, RNA tạo ra protein từ đó quy định các tính trạng của sinh vật. Khi cần tạo ra một protein, DNA sẽ tạo ra nhiều RNA với chức năng tổng hợp nên protein đó. Quá trình này sử dụng các phân tử tRNA (ARN vận chuyển) mang các axit amin đến phức hệ ribosome, nơi các phân tử rRNA (ARN ribosom) thực hiện gép nối các axit amin với nhau tạo thành chuỗi tiền protein.

Trong PPINs, người ta cho rằng các protein tương ứng với các gen không hoạt động (tức là:chưa được biểu hiện) sẽ không tương tác với các protein đối tác tiềm năng của chúng. Do đó, sản phẩm RNA do gen tạo ra thường được sử dụng như là một đại diện cho hoạt động của gen và điều này có liên quan tới việc kích hoạt các hệ thống phân tử trong PPIN và các quá trình sinh lý và phát triển. Nói cách khác, giá trị biểu hiện gen được tính thông qua lượng RNA được sinh ra, khi một gen biểu hiện tức là giá trị của nó vượt qua một ngưỡng được quy định thì protein tương ứng với gen đó trong mạng PPINs sẽ hoạt động.

Như vậy, các protein trong mạng PPINs có hai trạng thái hoạt động và không hoạt động, sự ổn định của mạng sẽ phụ thuộc vào các trạng thái này và được tính dựa trên thuật toán Deterministic Simulated Annealing (DSA). Thuật toán DSA được áp dụng trong việc so sánh độ ổn định của các mạng khác nhau dựa trên năng lượng của chúng. Cụ thể, năng lượng trong

PPINs của ung thư cao hơn năng lượng trong PPINs của AD, dẫn đến mạng tương tác protein của ung thư kém ổn định hơn mạng tương tác protein của AD.

2 Dữ liệu và phương pháp

2.1 Dữ liệu

Dữ liệu biểu hiện gen và tương tác protein được lấy từ các tập dữ liệu Gene expression và PPIN.

2.1.1 Mạng tương tác protein (PPIN)

PPIN của người được lấy từ cơ sở dữ liệu phân tích tương tác protein ($PINA$, phiên bản tháng 10, 2011. Tài nguyên trực tuyến 1: Wu et al. 2009). PINA là một nền tảng tích hợp dữ liệu PPIN đã được trích xuất từ sáu cơ sở dữ liệu khác nhau: IntAct, MINT, BioGRID, Dip, HPRD và MIPS / MPact. Nó bao gồm tự tương tác, và tương tác giữa protein người và protein từ loài khác. Hơn nữa, gần đây nó đã được sử dụng trong các nghiên cứu tương tự khác (Xia et al.2011; Laakso và Hautaniemi 2010).

Ngoài mạng PINA, hai mạng PPIN cũng được bổ sung để đảm bảo thu được kết quả tương tự: Cơ sở dữ liệu tham khảo về protein người(HPRD, http://www.hprd.org/doad, phiên bản tháng 4 năm 2010) có chứa các cặp tương tác protein của người dựa trên bằng chứng thực nghiệm từ tài liệu và đã được sử dụng trong một số nghiên cứu (Teschendorff và Severini 2010; West et al. 2012); Human Integrated protein–protein interaction rEference (HIPPIE, http: //cbdm.mdc-berlin.de/tools/hippie/doad.php, phiên bản tháng 9 năm 2014) kết hợp bộ dữ liệu PPI người với biểu đồ score được chuẩn hóa, tích hợp dữ liệu từ HPRD, BioGRID, IntAct, MINT, Rual05, Lim06, Bell09, Stelzl05, DIP, BIND, Colland04, Lehner04, Albers05, MIPS, Venkatesan09, Kaltenbach07 và Nakayama02. Các tương tác đã được chọn từ PPIN này với score trên 0,73 để tự đảm bảo các cặp protein tương tác với nhau (Schaefer et al. 2012)

2.1.2 Tập dữ liệu biểu hiện gen

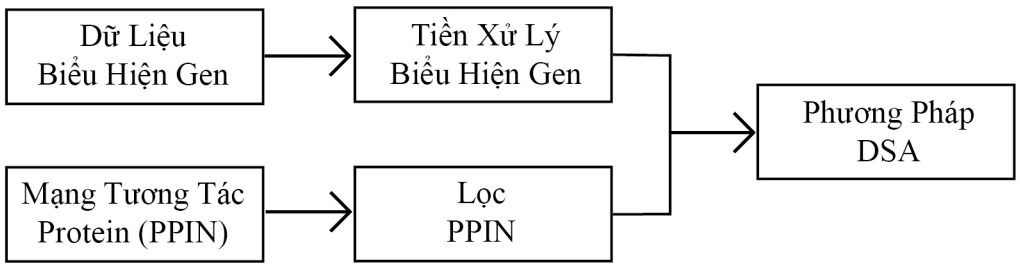
Hàng ngàn bộ dữ liệu biểu hiện gen có sẵn trong cơ sở dữ liệu được công bố, mỗi bộ chứa một mô tả về nguồn gốc y sinh tương ứng của mẫu, các quy trình phân tích và kết quả thử nghiệm về mặt biểu hiện (tức là: lượng RNA được tạo ra cho từng gen trong bộ gen).

Dữ liệu thô về biểu hiện gen (tệp CEL) cho Buồng trứng, Đại tràng, Gan được tải xuống từ Barcode human transcriptome repository (Gene Expression Barcode, http://barcode.luhs.org/), ngoài ra bộ dữ liệu SCZ và AD được tải xuống từ NCBI GEO omnibus (GEO, http://www.ncbi.nlm.nih.gov/geo) và Cơ sở dữ liệu bộ gen trực tuyến của Viện nghiên cứu y tế Stanley (SMRI, https://www.stanleygenomics.org: Tài nguyên trực tuyến 2). Mỗi tập dữ liệu tương ứng với một tập hợp các mẫu bệnh và mẫu đối chứng. Để phân tích, các trường hợp có quá ít các ca mắc bệnh / đối chứng (dưới 9) được loại bỏ và được lấy ra trong cùng một nền tảng (Affymetrix array GeneChip Human Genome U133 Plus 2.0), hiển thị thông tin về 23.945 gen người.

2.2 Phương pháp

Khi một đột biến sảy ra với một protein quan trọng mà có nhiều liên kết với các protein khác thì có thể gây ra sự thay đổi các quá trình sinh học. Một PPIN được lọc (Phần 2.2.1) và dữ liệu biểu hiện gen đã được xử lý trước và chuẩn hóa (Phần 2.2.3) cho ba trạng thái khác nhau

(ung thư, rối loạn bình thường và thần kinh) là đầu vào cho cách tiếp cận của chúng tôi (Phần 2.2.4). Sơ đồ của quy trình công việc được trình bày trong hình 1, trong đó tiền xử lý dữ liệu biểu hiện gen và lọc mạng tương tác protein được chia ra thành hai phần riêng biệt.



Hình 1: Biểu đồ phương pháp thực hiện

2.2.1 Lọc mạng tương tác Protein