異なる実験条件で得られたプロテオームデータの統合ネットワーク解析とエンリッチメント解析

Integrated network analysis and enrichment analysis of proteome data obtained under different experimental conditions

西崎愛花[1],荒木令江[2],河野信[1,3]

1.北里大学大学院未来工学研究科,2.熊本大学大学院生命科学研究部(医),3.情報・システム研究機構 ライフサイエンス統合データベースセンター

1 研究背景・目的

jPOST をはじめとする ProteomeXchange リポジトリから様々な実験で得られたプロテオームデータがオープンデータとして公開されている。プロテオームの解析を行う際には単一プロジェクトのデータだけでなく、異なる施設や異なるサンプルから得られたデータを統合して比較できると有用である。しかしながら、これらの異なる実験条件で得られたデータを単純に比較することは難しい。メタボロームの分野では、異なる実験条件のデータをネットワークを介して解析する試みがなされており[Matsuta, R., *et al. BMC Bioinformatics* **23**, 508 (2022)]、この手法をプロテオームデータにも適用した。
本研究では、異なる実験条件で得られたプロテオームデータを統合して解析するためにネットワークの作成を行う。また、構築されたネットワークが妥当であるか検証を行う。

2 研究方法

データセット

quantMS [Dai, C., *et al. Nature Methods* **21**, 1603–1607 (2024)]で再解析されているラベル化定量のデータで、サンプル間の比率が計算されている 36プロジェクト、2370 実験条件を用いた。また、CPTAC [Edwards, NJ., *et al. Journal of Proteome Research* **14**, 2707–2713 (2015)]の49プロジェクト、5260実験条件のデータを用いた。

タンパク質変動の有意性評価に向けたクオリティチェックの実施

quantMSおよびCPTACの全プロジェクトにおけるタンパク質の$\log_2$ fold changeの値を用い、プロジェクトごとに箱ひげ図でプロットすることで変動の分布を視覚化した（図1）。
平均から3倍の標準偏差を超えた値を外れ値として除外した。さらに、全データにおける95%信頼区間を算出し、各プロジェクトにおけるデータのばらつきと信頼性を評価した。これにより、タンパク質の変化量が統計的に有意であると判断するための基準が明確になった。

定量プロテオームデータに基づく実験群間ネットワークの構築

はじめに、quantMSで再解析されたデータのみを用い、2実験条件間の比較(=実験群) において、タンパク質変化量が2倍以上、かつ多重検定補正後のp値が0.01以下のタンパク質を取得した。各実験群間で検出されたタンパク質の変動についてクロス集計を作成して、タンパク質の増減挙動についてのオッズ比ならびにカイ二乗検定で p 値を計算した。カイ二乗検定で有意（有意水準 0.05 を Bonferroni 法で多重検定補正した p 値 $< 1.78 \times 10^{-8}$）であった実験群の組み合わせから、Cytoscapeを用いてネットワークを構築した（図2）。さらにカイ二乗検定で有意であった実験群の組み合わせから、異なるプロジェクト由来の実験群の組み合わせかつエッジが2本以上のノードを抽出してネットワークを構築した（図3）。次に、quantMSデータとCPTACデータセットを統合し、ネットワークを構築した。この解析では、コントロール群との比較を行っている実験群のみを対象として、FDR (Benjamini-Hochberg法）を用いて有意な組み合わせを抽出した（図4）。

カイ二乗検定

$$\chi^2 = \Sigma \frac{(O-E)^2}{E}$$

O = 頻度の観測値
E = 帰無仮説の下における頻度の期待値（理論値）

オッズ比計算式

$$odds\ ratio = \frac{m_{1,1} * m_{2,2}}{m_{1,2} * m_{2,1}}$$

クラスタリング解析と臨床情報との関連性評価

quantMSとCPTACの統合ネットワークにおいて、quantMS由来のハブノードごとに、接続されたCPTAC由来の実験群からタンパク質の発現量の値を用いてクラスタリングを行い、ヒートマップを作成した（図5）。
また、ネットワーク内の特徴を明らかにするため、CPTACのclinical dataの各項目に対してカイ二乗検定を実施し、統計的関連性を評価した。
クラスタリングされたタンパク質について、DAVID [https://david.ncifcrf.gov/tools.jsp]を用いてエンリッチメント解析を行った。
本研究で使用したデータセットの元論文ではがんのサブタイプが報告されているものも存在するが、本解析に用いたclinical dataファイルにはサブタイプ情報が含まれていなかったため、今回の解析ではサブタイプ別の検討は行わなかった。

3 結果

①各プロジェクトにおけるタンパク質変動の分布

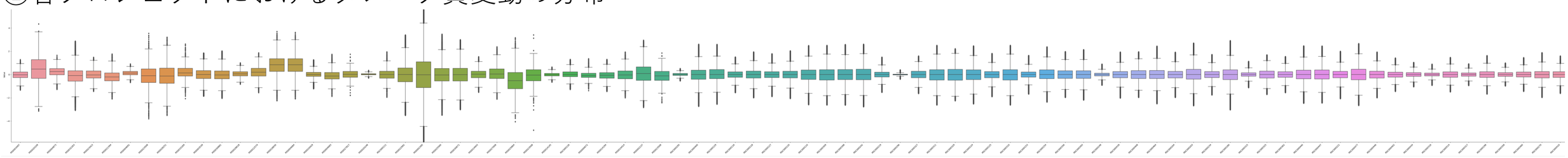

図1　quantMSとCPTACの各プロジェクトにおけるタンパク質変動(log2FC)の箱ひげ図

95%信頼区間が -1.0279～0.9171であった。よって、2倍以上増減しているものを有意なタンパク質変動とみなした。

②quantMSから構築したネットワーク

quantMSのネットワークは **1705** ノード、
有意水準を満たしたエッジの本数は **7.2万本**
で、**7個** のサブネットワークに分類された。

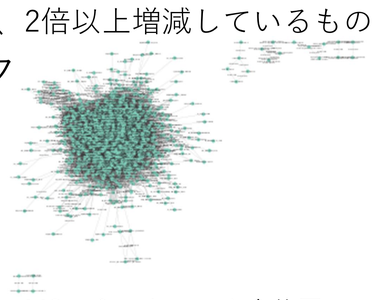

図2　quantMSのネットワーク全体図

quantMSの異なるプロジェクトの実験群の組み合わせから作成したネットワークは **41** ノード、有意水準を満たしたエッジの本数は **77本** であった。
ノードの色は実験群のクラスターに対応している。

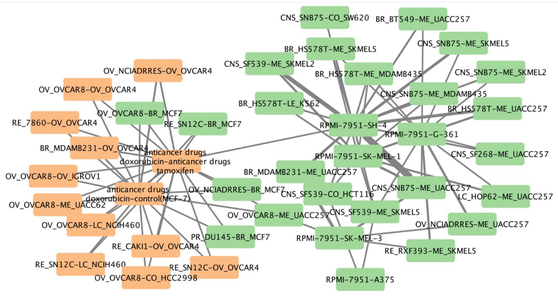

図3　発現挙動が似ている実験条件のうち、異なるプロジェクトの組み合わせかつエッジが2本以上からなるネットワーク

③quantMSとCPTACの統合ネットワーク

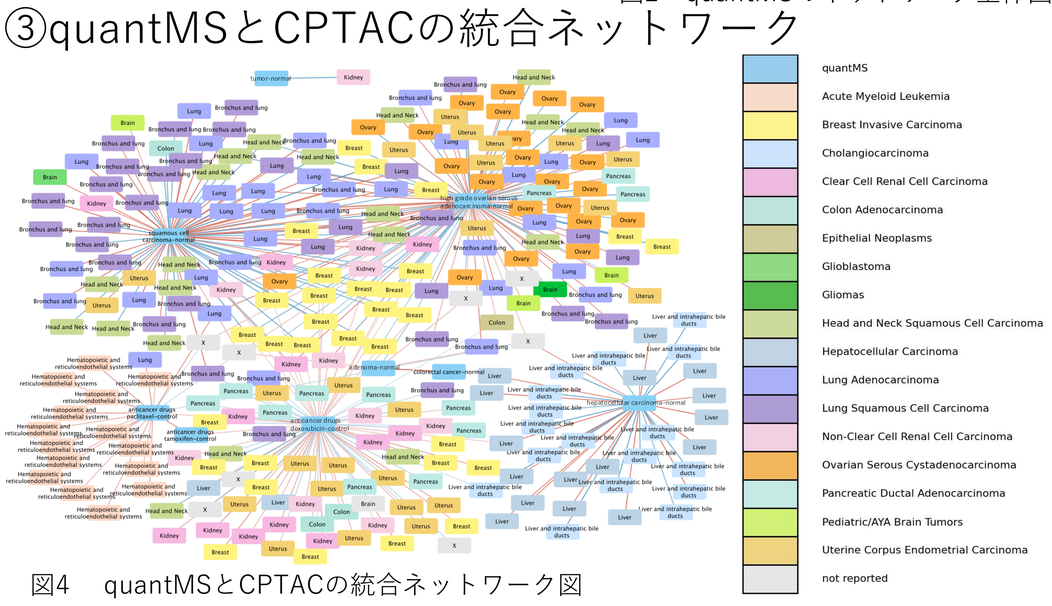

図4　quantMSとCPTACの統合ネットワーク図

統合ネットワークは**275** ノード、有意水準を満たすエッジは **345** 本だった。

④ エンリッチメント解析

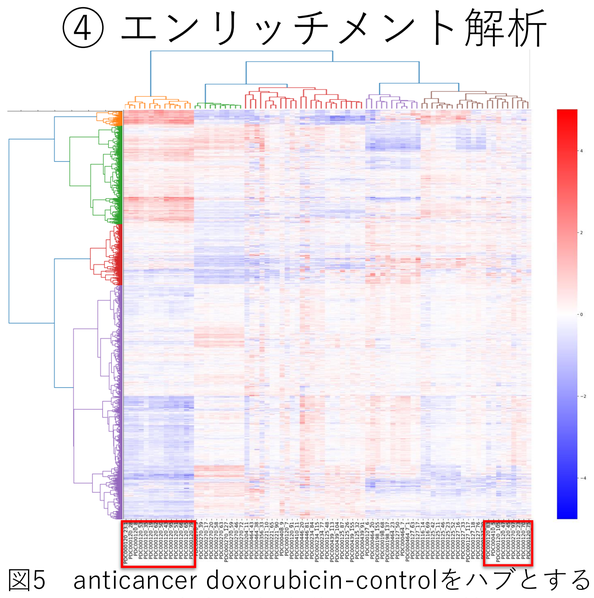

図5　anticancer doxorubicin-controlをハブとするCPTAC由来の実験群とタンパク質発現のヒートマップ

本研究では、 quantMS由来の**anticancer doxorubicin-control** のハブに接続された乳がん患者（図5の赤枠領域）に着目した。この患者群は2つの主要なクラスターに分割され、臨床データの解析により、Tumor Stageとの有意な相関が認められた（p値 = 0.0196）。
特に、左側のクラスターでは、がんステージがより進行している傾向が観察された。また、このクラスターで特に増加しているタンパク質を対象にエンリッチメント解析を行ったところ、**コレステロール代謝パスウェイ**が有意に関係していた（$p = 2.0 \times 10^{-6}$）。先行研究ではLuminal Aの乳がんにおいて、がん組織と非がん組織の比較で同パスウェイの活性化が報告されている [Meimei, Z., *et al., Cancer Medicine* **13**, e70470 (2024)]。
さらに、同じクラスターのGO解析でも脂質複合体の形成や組織修復、細胞外マトリックス再構築に関連する項目が上位に挙がっており、パスウェイ解析結果と一致した。

4 まとめ・展望

- 公開プロテオーム定量データから、タンパク質の発現挙動が似ている実験条件のネットワークを作成した。
- クラスター解析とエンリッチメント解析の結果、乳がん細胞の実験群が2つのクラスターに分かれ、がんのステージに有意に関係していることがわかった。
- がんのステージが進行しているクラスターのコレステロール代謝パスウェイの活性化は、先行研究で報告された乳がん(Luminal A)の知見と一致していた。
- 今後の展望として、今回作成されたネットワークから既知の疾患等の関係性について再発見することが可能であるか検証することに加え、ネットワーク上で関連が見られた、未知の関係について推定を行う。
- 今回はラベル化定量のデータセットのみを使用したが、ラベルフリー定量のデータも公開されているので、これらを加えてネットワークを構築し、分析を行う予定である。

本研究は、科学技術振興機構（JST）統合化推進プログラム　「創発的再解析のためのメタボローム統合データベース」（JPMJND2403）の支援のもと実施されました