博士論文（要約）

遺伝子発現パターンの解釈性向上を目指した
グラフ畳み込みネットワークの実装と検証
―深層学習の悪性腫瘍フェノタイプ分類への適用―

早　川　仁

ゲノムやトランスクリプトームを始めとする生体情報は、まとめてオミクスデータと呼ばれている。疾患と関わりのあるオミクスデータの情報は診療に応用されつつある。一方、まだ疾患との関連の意義が明確でないオミクスデータも多い。研究目的で測定されたオミクスデータは数多くがデータベースで公開されており、それらの二次利用による研究にも科学的な価値をもたらすことが期待されている。また、研究によって明らかになった知識は Kyoto Encyclopedia Genes and Genomes (KEGG)や Gene Ontology などのデータベースにも集約されている。これらの知識データベースに収められている情報には、例えば、生体の分子の相互作用や化学反応に関するパスウェイについてまとめられたものがある。ここにまとめられた情報は遺伝子とその機能を解析するためにも利用されている。細胞内では何らかの機能に関連した遺伝子は、関係の深い遺伝子同士がまとめて発現が変動することが知られている。オミクスデータの中でも転写産物の分析を行うことは細胞の遺伝子の発現パターンを把握するのに有用である。そのために用いられる手法として代表的なものに gene set enrichment analysis (GSEA)がある。GSEA は複数の遺伝子セットをまとめて解析することで、単一の遺伝子ごとに解析するよりも、より高い感度で機能に関連した遺伝子の発現量の変動を捉えることができる上に、遺伝子セットの意味のある変動を検出しやすいと考えられている。この解析を行うために使用する遺伝子セットには、生物学的な機能単位でまとめられたものを用いることが多いが、KEGG のような知識データベースに集約されたものが利用される。遺伝子セットで KEGG のデータベースを用いることで、GSEA によって特定のフェノタイプに特異的に高発現しているパスウェイを推定することができる。しかし、GSEA には批判もある。GSEA は変数とフェノタイプの一方向的な関係性しか検出できず、逆相関したり、複雑な関連がある場合は検出できない。そのため、GSEA に代わる手法がいくつも提案されている。本研究では近年発達してきた手法である深層学習を用いて GSEA で行っているようなパスウェイの分析ができないかを検証する。深層学習の一手法であるグラフ畳み込みニューラルネットワーク(GCN)は、グラフネットワークを利用して、あるノードに対して関連の強いノードの要素を畳み込む処理を行い、ノードとその周辺のノードから特徴量を取り出すことができる。この技術を用いることでグラフネットワークに表現した遺伝子のパスウェイから特徴量を取り出してフェノタイプとの関連について学習させる。さらに、このネットワークを用いた深層学習モデルを解釈することで、フェノタイプを予測するのに寄与するパスウェイを選択できないかを検証する。本研究では、びまん性大細胞型Ｂ細胞リンパ腫(DLBCL)の遺伝子発現量から分類されるサブタイプである germinal-center B-cell-like (GCB)タイプと activated B-cell-like (ABC)タイプを研究対象とした。遺伝子発現量のデータベースより DLBCL のデータセット(GSE31312, GSE10846)を取得して、腫瘍組織の遺伝子発現量と、それに対応するサブタイプについて得た。前処理を行った後、GSE31312 には GCB が 227 サンプル、ABC が 199 サンプル含まれており、GSE10846 には GCB が 183 サンプル、ABC が 167 サンプル含まれた。GSE31312 を深層学習モデルの構築を行う訓練データセットとして、GSE10846 をモデルの検証を行うテス

トデータセットとした。GCN を構築するために、それぞれが一つの KEGG パスウェイに相当する 186 のグラフネットワークにパスウェイを表現した。このグラフネットワーク上のグラフ畳み込みを遺伝子発現量に対して適用して DLBCL のサブタイプの分類予測を行うモデルを、訓練データセットで学習させた。この深層学習モデルを用いた分類予測のパフォーマンスを他の深層学習モデル（多層パーセプトロン(MLP)モデル、GCN-MLP モデル）と比較した。さらに、分類モデルを解釈するために Shapley additive explanation (SHAP) を用いることで、深層学習モデルのネットワークの中で、パスウェイの出力に相当する部分が予測にどのように寄与しているかを分析して、重要な貢献度をもつパスウェイを選択した。選択されたパスウェイと、GSEA でフェノタイプ特異的にエンリッチされたパスウェイを比較した。さらに、GCN により選択された上位のパスウェイが下位のパスウェイよりも分類予測の性能に大きく寄与していることを検証するため、SHAP による寄与の順位に基づいてパスウェイを選択して、選択されたパスウェイの遺伝子発現量を用いて分類予測をするロジスティック回帰モデルでの分類性能を比較した。学習済みモデルによるテストデータセットでの予測性能は、GCN モデルでは accuracy が 0.903、precision が 0.972、recall が 0.820、F1 score が 0.890 であり、他のモデルよりも優れた accuracy を示した。SHAP を用いて GCN モデルの解釈を行うと、Glycosaminoglycan biosynthesis – keratan sulfate、Limonene and pinene degradation、Pantothenate and CoA biosynthesis が分類予測に大きく寄与するパスウェイとして選択された。これらのパスウェイは GSEA でもフェノタイプ特異的にエンリッチされたパスウェイであった。また、GSEA では上位にリストアップされないが、フェノタイプと強く関連する B cell receptor signaling pathway も GCN モデルでは分類予測に大きく寄与していた。このことから、DLBCL のデータセットでは、GCN モデルはフェノタイプとの関連のあるパスウェイを選択して、さらに、GSEA で検出しにくいフェのタイプとの関係性をもつパスウェイの選択を行っている可能性が示唆された。また、GCN モデルで分類予測に大きく寄与するパスウェイに含まれる遺伝子発現量を用いたロジスティック回帰モデルでは accuracy 0.931、precision 0.933、recall 0.922、F1 score 0.928 の分類成績を示していた。GCN モデルでの分類予測に大きく寄与するパスウェイに含まれる遺伝子発現量を用いたロジスティック回帰モデルの分類成績は、寄与が小さいパスウェイに含まれる遺伝子発現量を用いたロジスティック回帰モデルの分類成績よりも良い傾向にあった。本研究の結果からは、GCN モデルを用いて遺伝子発現量から DLBCL のサブタイプの分類予測を行うことができた。さらに、そのモデルの解釈を行うことで、分類予測に寄与するパスウェイを選択することができる可能性が示唆された。