



## 未設定

1915077

水上和秀

情報基盤工学講座 指導教員 António Oliveira Nzinga René

## 要約

生命科学分野において、テキストマイニングを用いて、データベースに日々蓄積していく遺伝子データから、遺伝子・タンパク質間の関係性や相互作用を見出すことの重要性は依然として強い。本研究ではある生物が持つたんぱく質の共起関係を抽出し、それらの関係性を3Dグラフに可視化する。そして、得られた関係性について、さらなる分析を行うための支援を目的とする。

## キーワード：

遺伝子データベース、スクレイピング、テキストマイニング、共起ネットワーク、可視化

## 1はじめに

計算機の発展に伴い、年々増加しつつある大量かつ整理されていない文書データに対して、自然言語処理や情報検索技術等を用い、有用な情報を見い出すテキストマイニングが近年盛んになってきている。あるキーワードで検索し、検索結果に現れる文章から自然言語処理などを用いて必要な単語のみを抽出したり、得られた情報をデータベース化し、新たな情報提供を行えるソフトウェアを開発したりなど、様々な応用がある。生体・遺伝子情報を扱う生命科学分野においても、日々蓄積されていく文献・遺伝子データベースの中から特定の働きをもつ遺伝子名を検索したり、遺伝子間の関係性や相互作用を見い出したりするのにテキストマイニング技術が用いられている。

本研究では、遺伝子間関係を可視化することに重点を置いている。まず、遺伝子データベースであるKEGG(Kyoto Encyclopedia of Genes and Genomes)が提供する遺伝子データを用いて、ある生物が持つタンパク質どうしの共起頻度を計算する。そして、得られた共起頻度を重みとする隣接行列を作成し、共起ネットワークとしてタンパク質間の関係を描画する。

## -2 テキストマイニングと可視化-

## 2.1 スクレイピングとテキストマイニング

大量かつ整理されていない文書データから有用な情報を抽出するためには、元となる文書データを収集する必要があり、収集方法の1つとしてスクレイピングがある。スクレイピングとは、Webサイトから文章をプログラミングによって自動取得する方法であり、効率的にデータを収集できる。

テキストマイニングでは、自然言語処理を用いて、取得した文章を品詞レベルの単語に分解し、ある単語の出現頻度や、1つの単語に対して別のある単語が呼応して出現する共起頻度などを分析することで、有用な情報を抽出する。スクレイピングとテキストマイニングを組み合わせることで、効率的に情報抽出する仕組みを作ることができる。自然言語処理の1つである形態素解析の例を図1に示す。

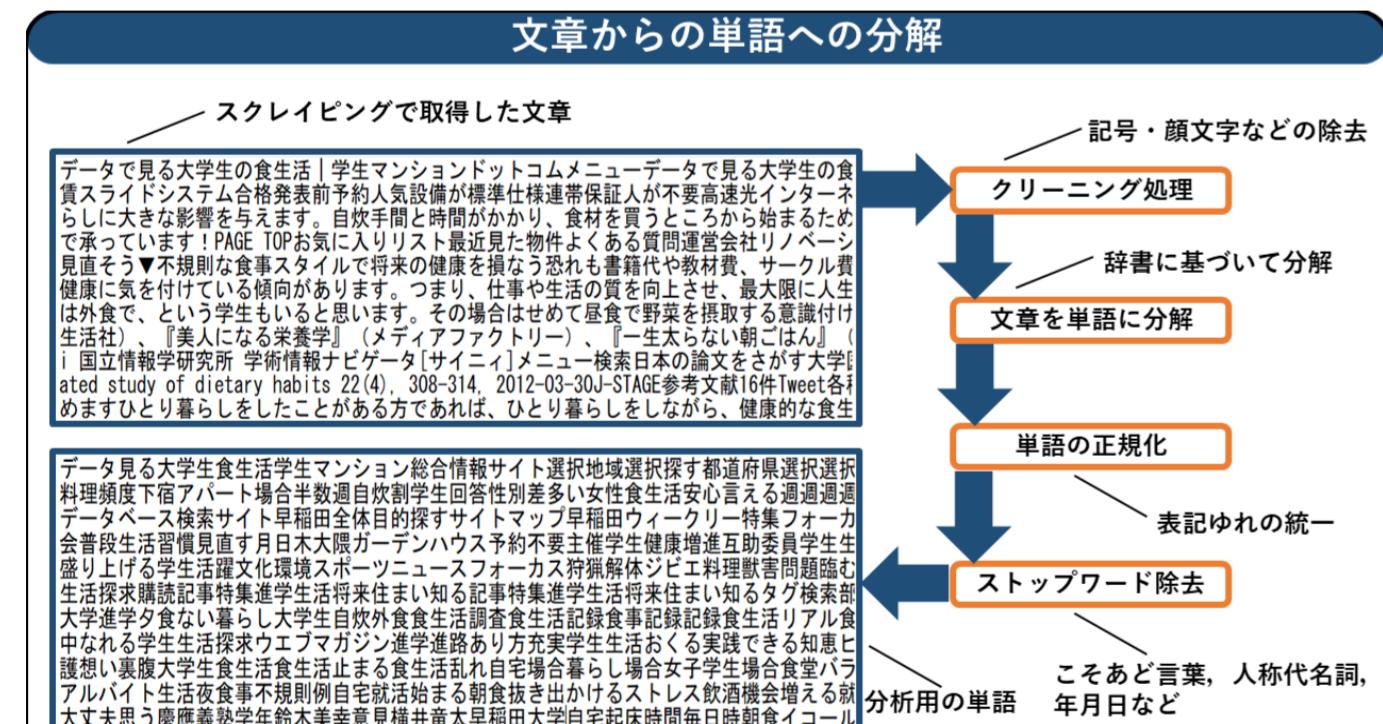


図1 形態素解析による文章の分解

## 2.2 可視化による情報抽出

テキストマイニングでは、単語どうしの共起頻度や、出現傾向の相関性といった分析結果を、頂点(ノード)と辺(エッジ)を持ったグラフを用いて可視化することが多い。ノードには抽出した単語が入り、関連のある単語同士がエッジで結ばれる構造となる。出現頻度や共起頻度の高さを、エッジ太さやノードの大きさで表すことによって、視覚的な分析を行うことができる。

3Dグラフ[1]では、あるワードを検索して表示されるWebサイトや、Twitterのツイートから文章をスクレイピングし、テキストマイニングによって共起頻度が高い単語同士が共起ネットワークで表現される。図2(左)が3Dグラフの例であり、共起頻度が高い単語ペアほど太いノードで表現されている。ユーザは検索ワードから複数の関連する単語を3Dグラフで見ることができ、アイデアの発想支援につなげることができる。

## -3 生命科学とテキストマイニング-

## 3.1 遺伝子間関係の可視化

生命科学分野におけるテキストマイニングでは、遺伝子データベースにある文献内の共起関係に着目した、タン

パク質間相互作用の関係性の可視化や遺伝子構造のネットワーク抽出などが挙げられる。

遺伝子データベースの1つにPubMedがある。PubMedとは生物・医学論文や妙録をデータベース化したものであり、今までに発表されてきた膨大な数の文献を調べることができる。PubMedにはMeSHと呼ばれる医学用語を階層構造に分類したシソーラスがあり、MeSH内の用語の共起性を用いて用語間の関係性を調べる研究が多くなってきた。

研究例として、MeSHから用語を抽出し、ビブリオメトリックの指標として、Dice係数の逆数を用いて2つの遺伝子間の非類似行列を求め、それをエッジ重みとして可視化したり[2]、サイトカインと疾患の関係性をMeSH用語の共起頻度を用い、相関行列を導出してクラスタリングを行ったものがある[3]。タンパク質の共起ネットワークの例を図2(右)に示す。

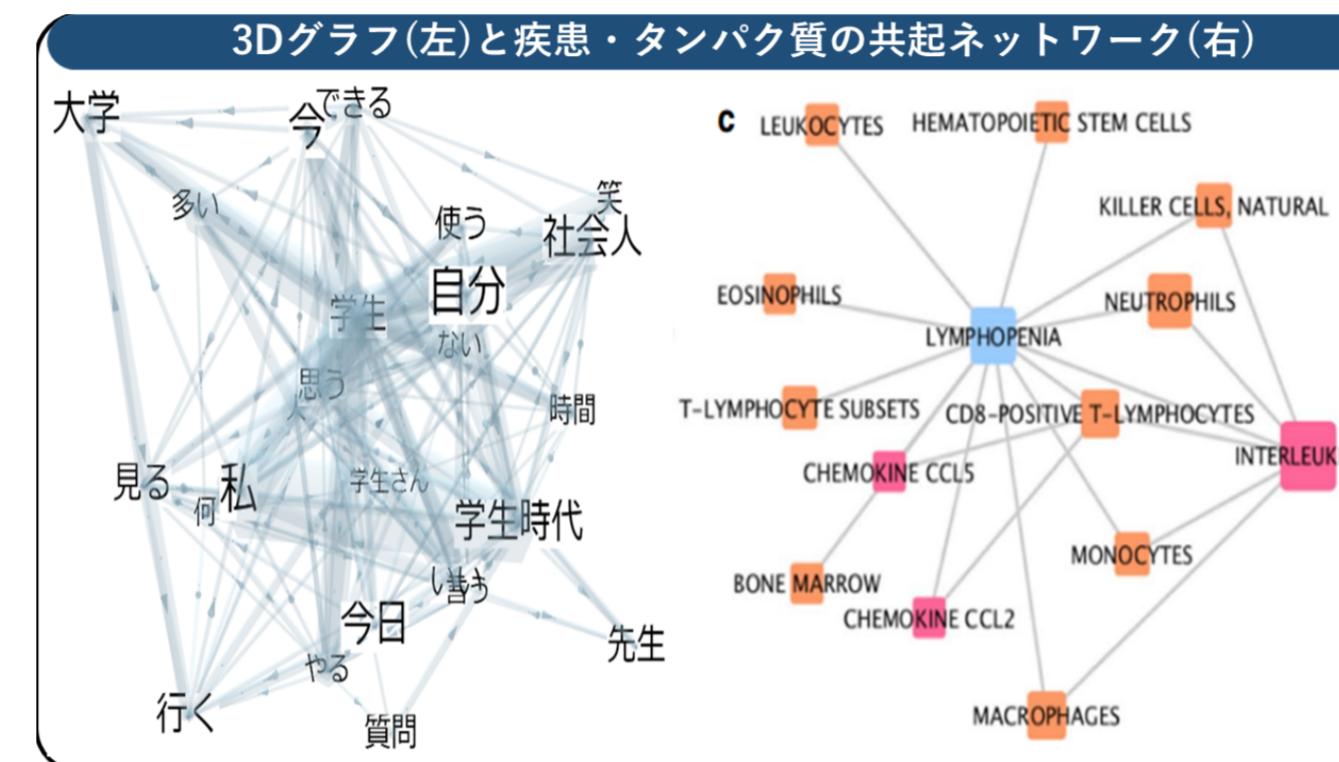


図2 3Dグラフとタンパク質の共起ネットワーク[2]

## 3.2 タンパク質間の共起性

遺伝子やタンパク質の関係性を調べる際に共起分析が良く用いられる。共起分析はテキストマイニングの研究ではよく用いられており、文章中のある単語に付随して出現する単語がある場合、その2つの単語には何らかの関連があるものとして分析する手法である。PubMedなどのデータベースを用い、論文中に出現する遺伝子やタンパク質の用語を抽出し、それらの共起性を分析する研究が多く行われている。

一方で、文章ではなく、生物のゲノム配列内やパスウェイに出現する遺伝子どうしの共起分析を行う研究も存在する。ゲノム配列の共起分析の例として、世界的にMRG(金属耐性遺伝子)が抗生物質耐性遺伝子(ARG)の増加に影響している傾向がみられることがから、可動遺伝因子(MGEs)のDNAに含まれるARGとMRGの共起性を解析した研究がある[4]。用いられている共起性の評価指標として、遭遇率と平均最小距離の2つがある(図3)。

遭遇率は1つのゲノム配列に対して、MRGから距離5000bp内にあるARGの数を [200bp, 100000bp](ステップ数:200bp)の範囲でカウントされる。MRG, ARGに属する遺伝子をそれぞれ  $M_i, A_i$  とし、この範囲における  $M_i$  の数を  $N_{(M_i)}$ 、1つのMRG遺伝子の前後5000bp内に含まれるARGの数を  $N_{(A_i)}$ 、とすると、 $K$  個のゲノムにおける  $M_i$  と  $A_i$  の遭遇率  $IoE_i(1)$  が得られる

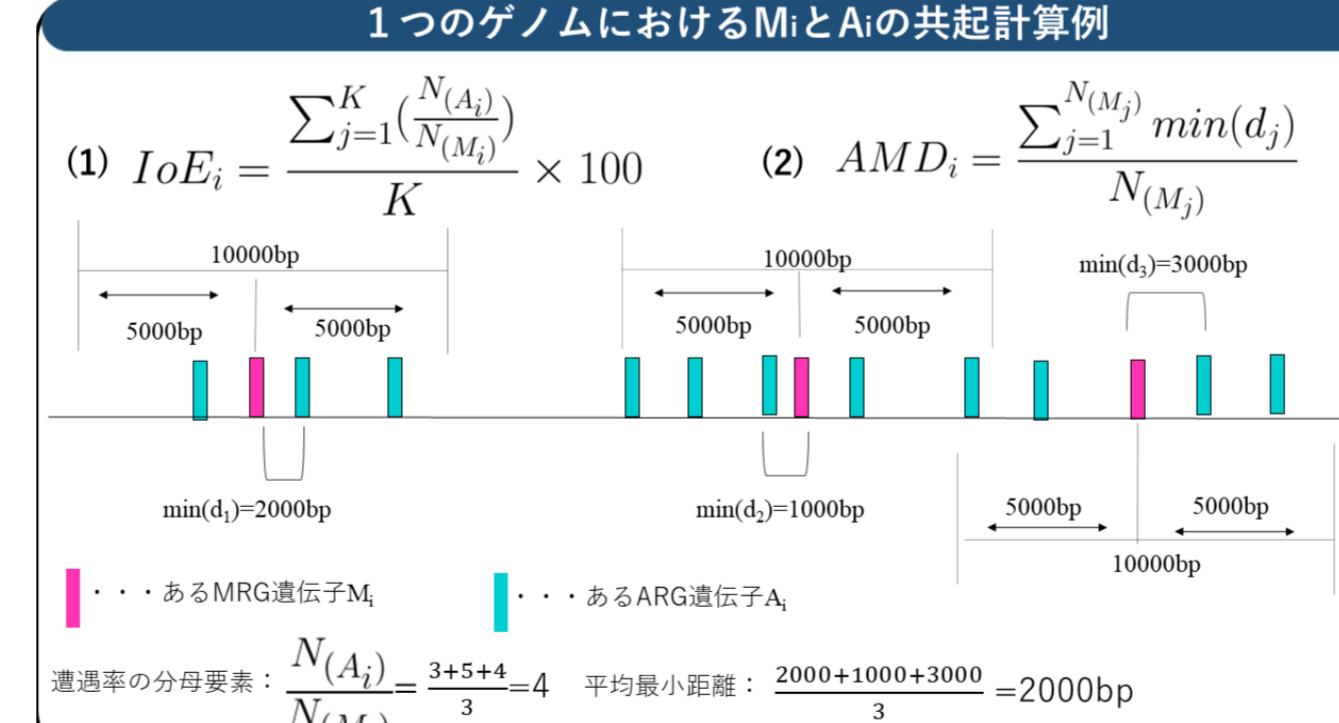


図3 共起頻度の計算法[4]

平均最小距離  $AMD_i$  は  $M_i$  の前後5000bp内に複数存在する  $A_i$  のうち、最も近くにある  $A_i$  と  $M_i$  の距離  $d_i$  における1ゲノム中の平均距離であり、(2)で与えられる。

## 4 提案手法

本研究では、パスウェイ内に同時出現する2つの遺伝子に着目して共起分析を行う。パスウェイとは代謝における酵素反応や、シグナル伝達といった遺伝子・タンパク質の相互作用を描画したマップのことであり、様々な遺伝子解析実験のデータを、専門家が手動で整理することで作成されている。

KEGGが提供するKEGG Pathwayから、特定の生物が持つ遺伝子とパスウェイの情報をKEGGのREST APIを用いて取得し、遺伝子とその遺伝子が含まれるパスウェイの対応表を作成する(図4)。

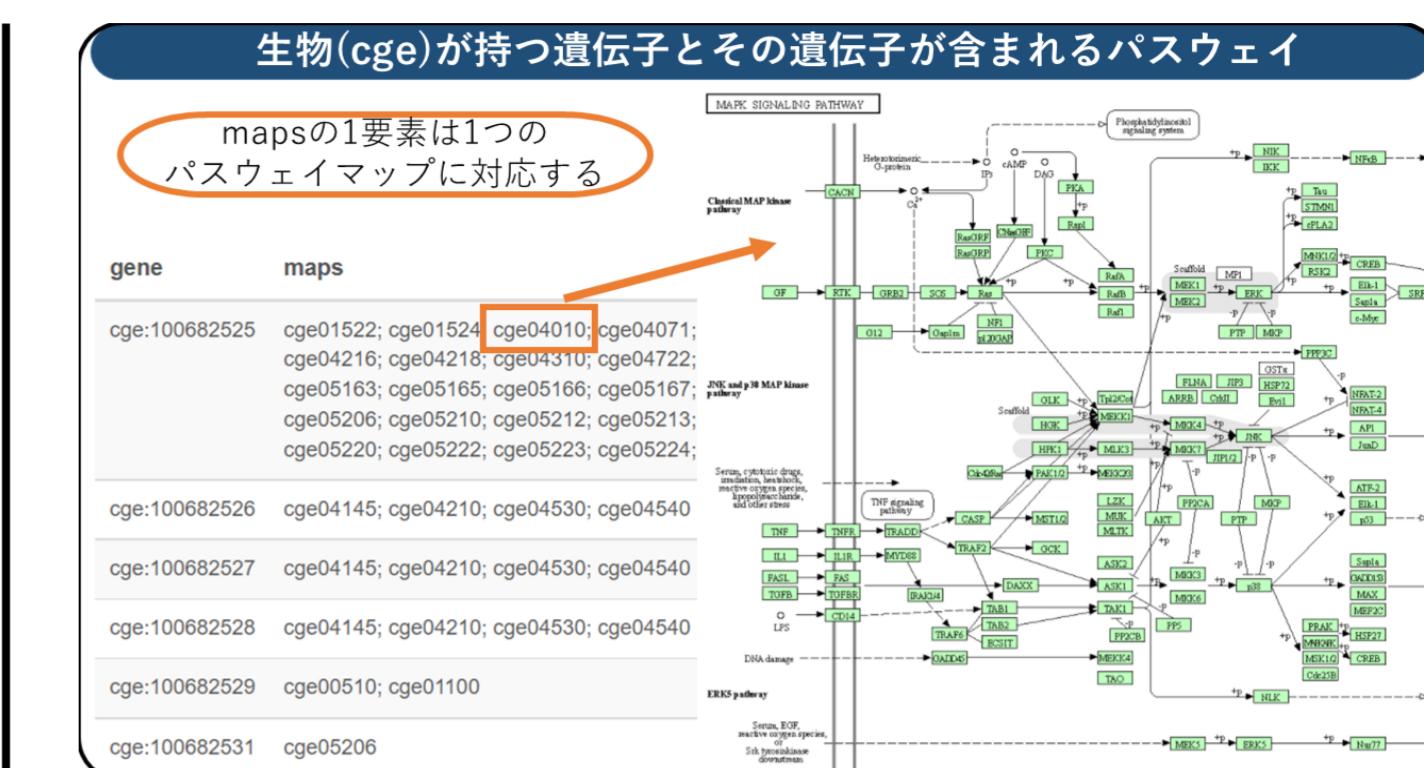


図4 対応表とパスウェイ

対応表内のある2つの遺伝子が同じパスウェイ内に出現する回数を計算し、共起頻度の重みとして隣接行列に表現する。そして、共起頻度が高い上位数十件(任意)のみに絞って3Dグラフに可視化する。最終的に出てきた関係について、どのような意味づけができるのか検討する(図5)。

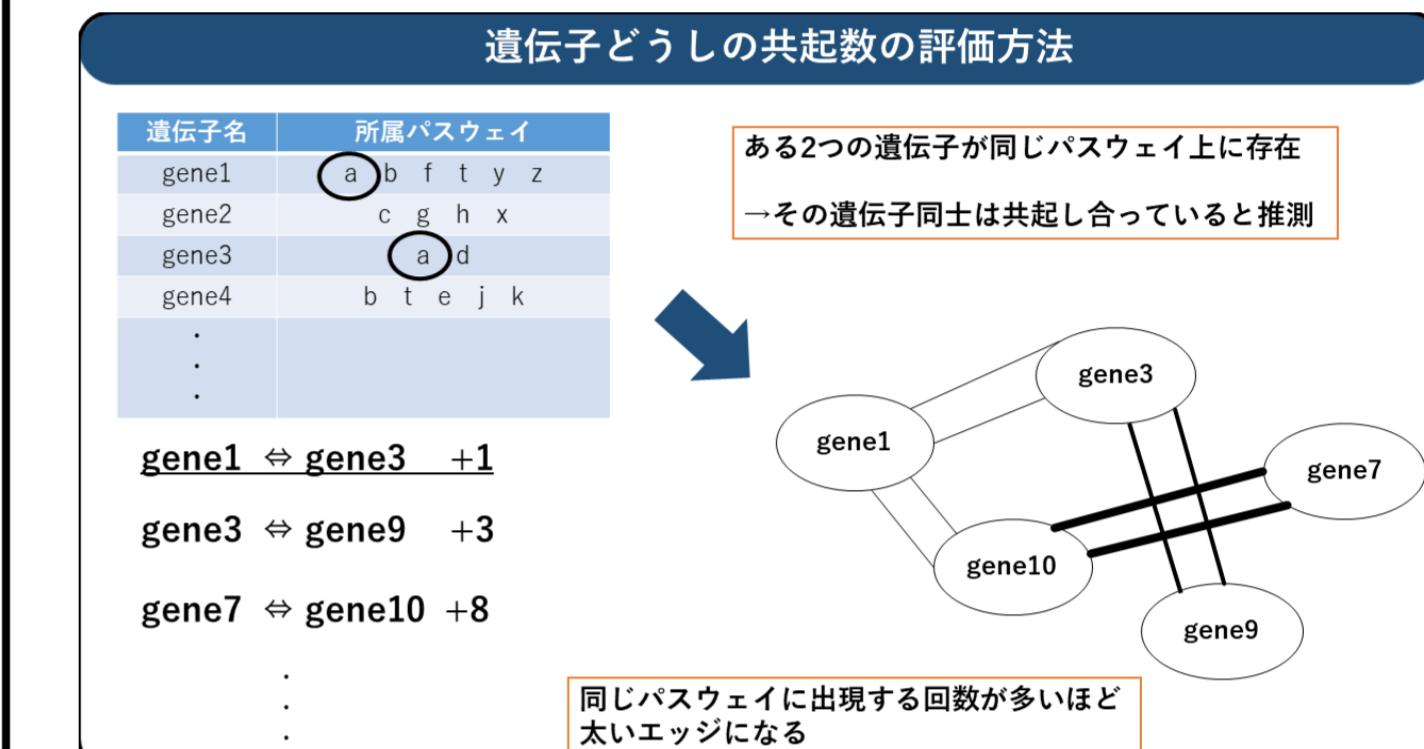


図5 パスウェイからの共起頻度評価方法

## 5 数値実験並びに考察

最初に、複数のヒト(接頭辞:hsa)の遺伝子とパスウェイを取得・整理し、対応表を作成した。次に、4章で述べた方法を用いて2つの遺伝子が同時に出現する頻度を計測し、図6(左)のような隣接行列を生成した。最後に、エッジの重みとなる隣接行列の値を、大きさに応じて10段階の階級に分類し、2段階ごとに色付けを行ったものを3Dグラフに出力した(図6(右))。

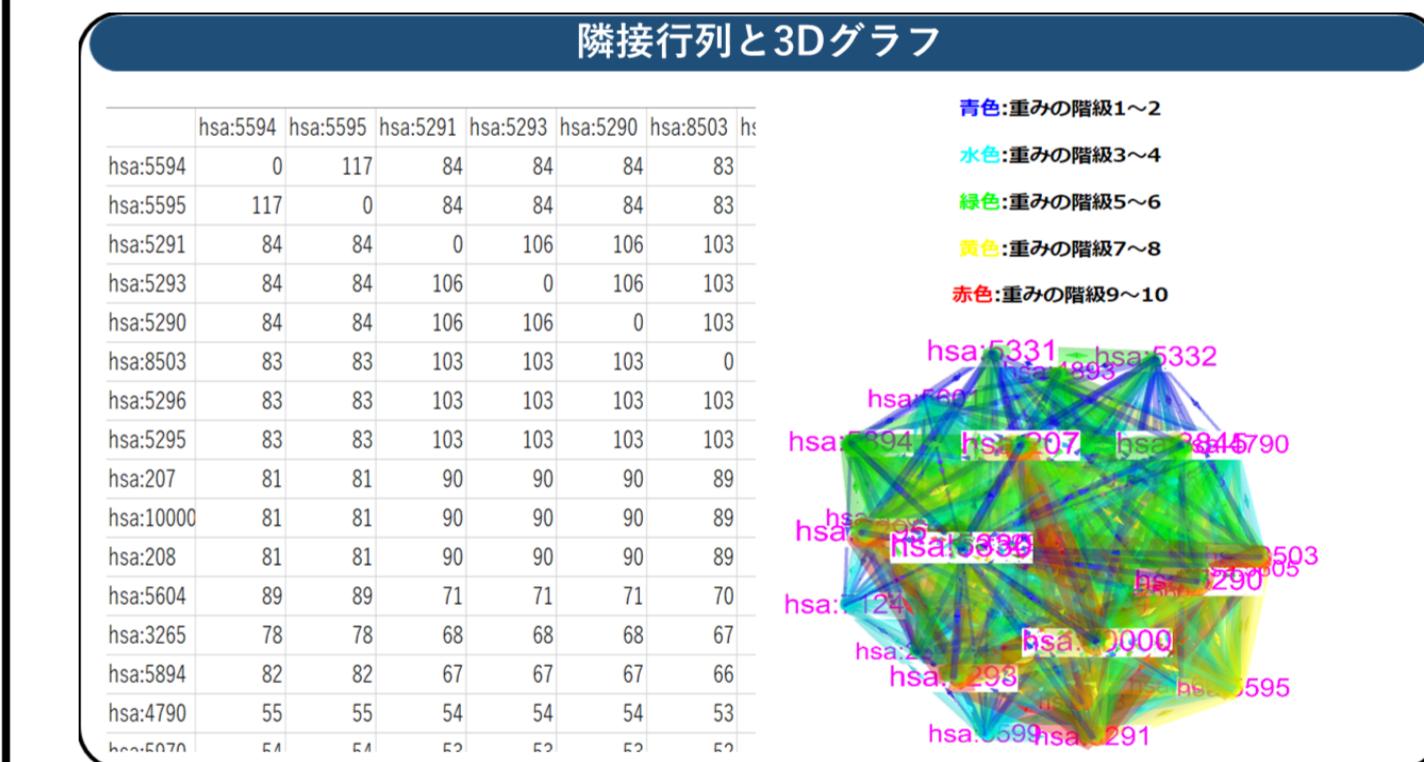


図6 ヒトが持つたんぱく質関係の可視化結果

最も強い赤色(共起度 99~110)のつながりとして5290,5291,5293,5295,5296,8503(hsaは省略)のネットワーク、207,208,10000のネットワークが得られた。前者の場合は、いずれのノードもホスファチジルイノシトール3-キナーゼ(PI3K)というシグナル伝達タンパク質のネットワーク、後者はセリン/スレオニンキナーゼ(Akt)の3種類のアイソフォームからなるネットワークであった。PI3K, Aktのどちらもがんに関連するものであり、ヒトのパスウェイの中でこの2つのネットワークが最も強く現れたことについて、さらなる分析につなげられると考えられる。

## 6 おわりに

本研究では、同じパスウェイに同時に出現するタンパク質に注目し、同時に出現する頻度をエッジの重みとして3Dグラフに出力した。今後の課題として、パスウェイの階層構造や1つのノードに対してつながっているノードの分析などを行って行きたいと考えている。

## 参考文献

- [1] 平松 楓也, "発想支援とジオプロセンシングのシームレスな統合に向けたQGISプラグインの開発", 富山県立大学学位論文 2020.
- [2] Stapley BJ, Benoit G, "Biobibliometrics: information retrieval and visualization from co-occurrences of gene names in Medline abstracts", Pacific Symposium on Biocomputing 2000.
- [3] Nophar Geifman, Anthony D. Whetton, "A consideration of publication-derived immune-related associations in Coronavirus and related lung damaging diseases", Journal of Translational Medicine 2020.
- [4] Li-Guan Li, Yu Xia, Tong Zhang, "Co-occurrence of antibiotic and metal resistance genes revealed in complete genome collection", The ISME Journal 2016.