遺伝子データベースからのテキストマイニングによる発展的分析支援のための遺伝子間関係の可視化

1815070 武藤克弥　　情報基盤工学講座　指導教員 奥原浩之

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され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行っ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，自然言語処理といった方法を用いて，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．

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．ここでは，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とは，生物・医学文献や遺伝子データベースから表記名・派生語が複数ある遺伝子名等を正確に識別したり，タンパク質間の相互作用の関係性やネットワーク状の遺伝子構造を抽出したりなどが挙げられる．とりわけ，関係性・構造抽出においては，視覚的な見やすさからグラフを用いて可視化されることが多い．遺伝子データベースの１つに**PubMed**がある．**PubMed**とは生物医学雑誌の論文や妙録を電子化したものであり，今までに発表されてきた膨大な数の文献を閲覧することができる．**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されてきた．

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りといったものがある**[3]**.

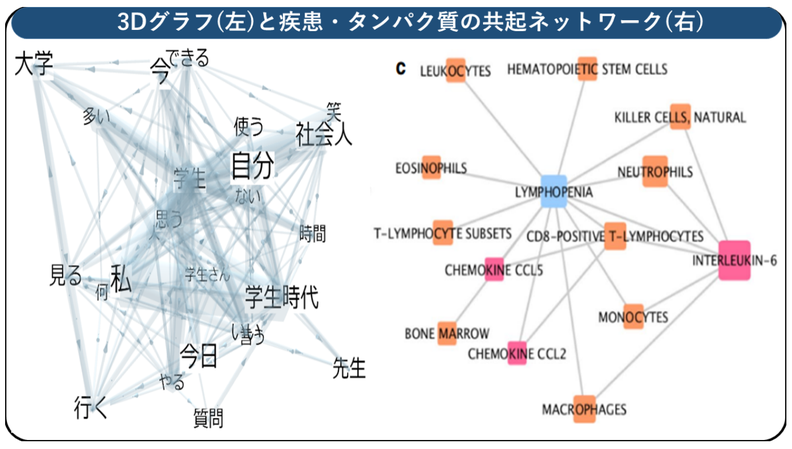

図1 **3D**グラフとタンパク質の共起ネットワーク**[2]**

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．生命科学分野においては**PubMed**などの論文の妙録を集めたデータベースを用い，論文中に出現する複数の遺伝子名やタンパク質名を抽出し，その共起性を分析する研究が数多く行われている．

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．**[4]**では，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**の共起性を解析している．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．

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

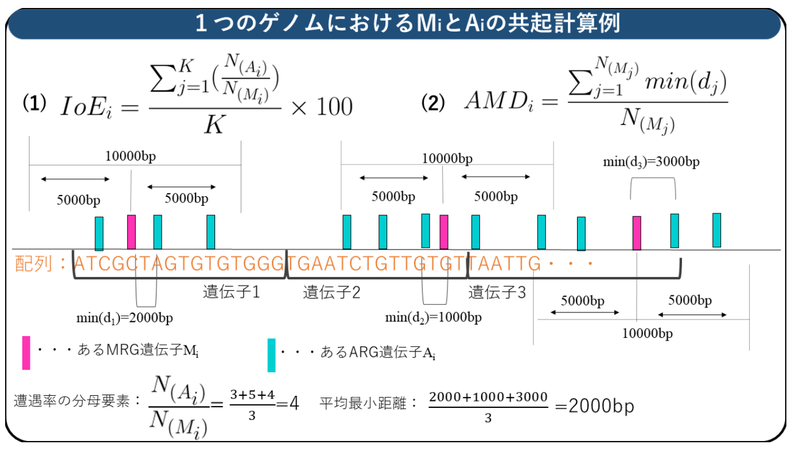

図2 共起頻度の計算法**[4]**

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．**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 API**を用いて取得し，遺伝子とその遺伝子を含むパスウェイの対応表を作成する．

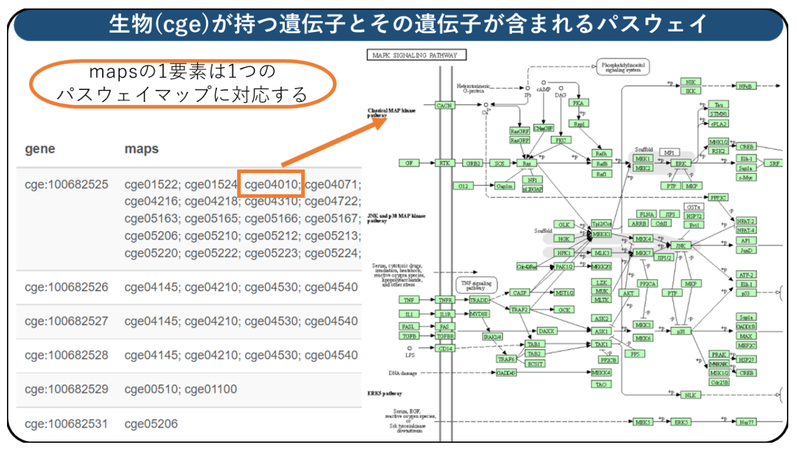

図3 対応表とパスウェイ

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測し，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．最終的には複数の生物においても，同様の方法で共起分析を行い，意味のある遺伝子関係が可視化されているか検証を行う．

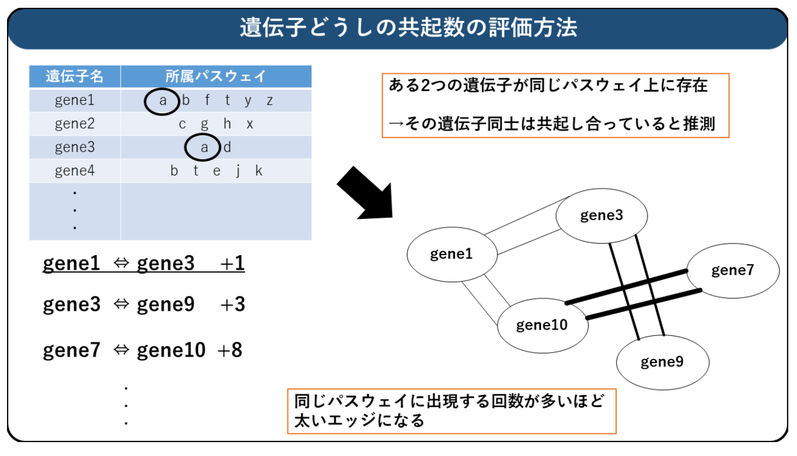

図4 パスウェイからの共起頻度評価方法

5 数値実験並びに考察

6 おわりに

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.